Automated Function Prediction – Biosapiens Joint Special Interest Group Meeting

Vienna, Austria July 19-20 2007

Meeting Program

AFP / Biosapiens 2007 Agenda

DAY 1	Thu, Jul 19, 07	
09:00-09:10		Opening remarks
09:10-09:50	**David Jones**	**Predicting functions of human proteins using patterns of native disorder**
09:50-10:10	Nils Weinhold, Oliver Sander, Francisco S. Domingues, Thomas Lengauer, Ingolf Sommer	Molecular Function Prediction based on Local Function Conservation in Sequence and Structure Space
10:10-10:30	Shula Shazman, Yael Mandel-Gutfreund	Predicting RNA-binding Proteins from their Three Dimensional Structure
10:30-11:00		Coffee Break
11:00-11:40	**Russ Altman**	**Modeling Dynamics of Ca2+ Binding Sites: A Glimpse into 4-Dimensional Space**
11:40-12:00	Oliver Redfern, Tim Dallman and Christine Orengo.	FLORA: A novel method for predicting function from structure
12:00-2:00		Lunch & posters
2:00-2:40	**Alfonso Valencia**	**The Genome-wide annotation schema in the BioSapiens-ENCODE collaboration**
2:40-3:00	Michal Linial, Noam Kaplan and Noa Morpurgo	Functional Prediction of Abandoned Short Sequences: Toxin-like Peptides from Insect and Mammalian Genomes
3:00-3:20	Anika Joecker *et al* / Nicolas Gagnière *et al* / James Bradford *et al*	Lightning talks and discussion: progress in genomic annotation
3:30-4:00		Coffee Break
4:00-4:20	Gaurav Pandey and Vipin Kumar	Incorporating Functional Inter-Relationships Into Algorithms for Protein Function Prediction
4:40-5:00	Mark Wass and Michael Sternberg	ConFunc: Feature derived Profiles for Functional Annotation

5:00-5:40	**Shoshana Wodak**	**Functional annotations of protein 3D structures with the GENEFUN meta function assignment resource**
5:40-6:00	Benoit Dessaily, Mark Lensink, Christine Orengo, Shoshana Wodak	FunSite: a gold standard dataset of ligand-binding sites in protein structures
6:00-6:20	Marcin von Grotthuss, Dariusz Plewczynski, Leszek Rychlewski	3D2EC: predicting enzyme function from structure

DAY 2	Fri, Jul 20, 07	
09:00-09:10		Opening remarks
09:10-09:50	**Janet Thornton**	**The Evolution of Enzyme Mechanisms and Functional Diversity**
09:50-10:10	Kieran O'Neill et al	Ontodas: integrating DAS with ontology based queries
10:10-10:30	Troy Hawkins, Meghan Chitale and Daisuke Kihara	Enriching functional knowledge in proteomics data using high throughput PFP prediction
10:30-11:00		Coffee Break
11:00-11:40	**Ewan Birney**	**ENFIN: computers meet experiments.**
11:40-12:00	Gabriele Schweikert, Georg Zeller, Alexander Zien *et al.*	mGene: a novel discriminative ab-initio gene finding system
12:00-1:00		Lunch
1:00-1:20	**Steven Brenner**	**Scaling SIFTER to Annotate larger, Functionally Diverse Protein Families**
1:20-1:50	Richard Albang *et al* / Sinan Sarac *et al* / Alexis Rodriguez *et al*	Lightning talks & discussion: emerging methods in function prediction
1:50-2:30	**Frederick Roth**	**A Critical Assessment of M. musculus Gene Function Prediction Using Integrated Genomic Evidence**
2:30-3:30	**Anna Tramontano**	**Panel Discussion: assessment of function prediction: lessons learned**
3:30-4:00		Coffee Break

DAY 2	Fri, Jul 20, 07	
4:10-4:30	Inkyung Jung, Jungsul Lee, Chulhee choi and Dongsup Kim	A novel functional annotation algorithm, combining protein similarity profile with a measurement to assess profile similarity
4:30-4:50	Annalisa Marsico, Andreas Henschel, Gihan Dawelbait, Christof Winter, Anne Tuukkanen and Michael Schroeder.	Conserved hydrogen bond patterns reveal structural and functional motifs in transmembrane protein regions
4:50-5:10	Andreas Prlic	DAS
5:10-5:50	**Rob Russell**	**TBA**
5:50-6:00		Closing remarks

PLENARY TALKS

Modeling Dynamics of Ca^{2+} Binding Sites: A Glimpse into 4-Dimensional Space

Dariya Glazer, Randy Radmer, Russ B. Altman*
Departments of Genetics & Bioengineering, Stanford University, Stanford, CA 94305-5120, USA
*To whom the correspondence should be addressed: russ.altman@ stanford.edu

1. INTRODUCTION

The development of high-throughput structural determination methods allows scientists to solve structures of proteins without a known function. Consequently, the Protein Data Bank (PDB) now contains an increasing number of entries without any assigned function. Computational function-prediction methods can assist in providing initial functional annotations that can reduce the number and types of assays required experimentally to confirm a protein function. Several bioinformatic approaches have been employed to address this challenge, assigning functions based on sequence and structure similarities and homology. Structure-based approaches may sometimes fail because the molecular structures solved by X-ray crystallography represent only a single, potentially artefactual, conformation. Therefore, there is a potential for using dynamic information to improve the function detection ability of these methods.

Molecular Dynamics (MD) techniques can simulate inter and intramolecular behavior over time (1). Simulation allows the molecule to relax by removing the crystallographic packing constraints, thus sampling a wider range of physiological conformations.

In the PDB, there sometimes exist pairs of structures for the same molecule, one of which manifests a function while the other does not. For example, there are proteins whose structures have been solved both with (holo) and without (apo) Ca^{2+} in the crystallization solution. We chose 5 such molecules to test whether conformations obtained from MD simulations could improve our ability to recognize Ca^{2+} binding function for these molecules. The initial papers describing these molecules have also reported differences between static Ca^{2+}-bound and unbound forms (1B9A and 1B8C, 1K96 and 1K8U, 1I40 and 1MJW, 3DNI and 1DNK, 1K94 and 1F4Q). We used the physics-based molecular dynamics software suite, GROMACS, to simulate behavior of both holo and apo structures of these 5 molecules explicitly solvated in water over time (usually 1-10 ns) (2).

FEATURE is a tool that scans 3-dimensional (3D) environments to produce a statistical model of key physicochemical characteristics found preferentially at the site of interest (3). We have created and validated Ca^{2+}-binding models before (3, 4). We used one such FEATURE model to evaluate structures generated by the MD simulations at various time points as a computational assay for the presence of Ca^{2+} binding sites. A 3D grid spaced at one angstrom and encompassing the whole protein allowed scoring of microenvironments centered at each grid point throughout the structure.

Through coupling MD simulations with the FEATURE scoring algorithm, we observed the appearance, persistence, and disappearance of the conformations that may accommodate Ca^{2+} binding (See Figure 1,2). At a given grid point, a score of 50 or greater using the current Ca^{2+} binding FEATURE model identifies a potential Ca^{2+} binding site. Our results indicate that MD sampling of structure indeed enhances the ability to assign putative function to a given structure. As expected, most static holo PDB structures scored above the FEATURE model threshold. MD simulations of these structures indicated that these proteins could achieve even more favorable Ca^{2+} binding conformations over time (See Figure 1a, Figure 2a,b,c,d). Some of the initial static apo PDB structures scored below the FEATURE model threshold, and therefore would not normally be annotated as Ca^{2+} binding proteins. However, high scoring Ca^{2+} binding conformations appeared at several instances over the course of the simulations (Figure 1b). Thus, our results indicate that MD simulation may explore structural diversity sufficiently to allow machine-learning based algorithms more accurately to identify function.

3. REFERENCES

1. Levitt, M. and R. Sharon, *Accurate Simulation of Protein Dynamics in Solution.* PNAS, 1988. **85**: p. 7557-7561.

2. Lindahl, E., B. Hess, and D.v.d. Spoel, *GROMACS 3.0: A Package For Molecular Simulation And Trajectory Analysis*. J Mol Modeling, 2001. **7**: p. 306-317.
3. Wei, L. and R.B. Altman, *Recognizing Protein Binding Sites Using Statistical Descritions Of Their 3D Environments*, in *PSB: Pacific Symposium of Biocomputing*. 1998: Maui, HI. p. 497-508.
4. Bagley, S.C. and R.B. Altman, *Characterizing the Microenvironment Surrounding Protein Sites*. Protein Science, 1994(4): p. 622-635.
5. Humphrey, W., A. Dalke, and K. Schulten, *VMD - Visual Molecular Dynamics*. J. Molec. Graphics, 1996. **14**(1): p. 33-38.

2. FIGURES

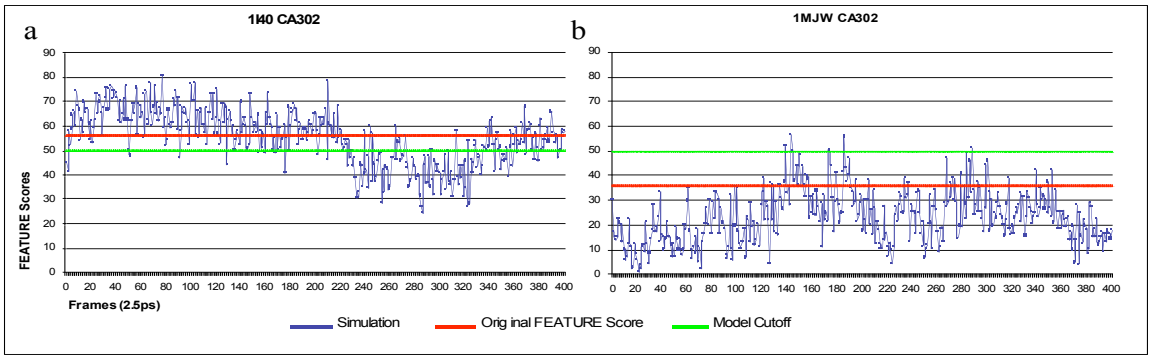


Figure 1. 1ns MD simulations were created for a) PDB_ID:1I40 and b) PDB_ID:1MJW, a holo-apo pair of static structures. Frames were extracted every 2.5ps, and scored using Ca^{2+} binding FEATURE model. In the holo form (left), the presence of Ca^{2+} leads to a high initial score (red) which is substantially maintained through the simulation. In the apo form (right), the absence of Ca^{2+} leads to a low initial score, but the simulation samples several conformations (frames 140-200) achieving high Ca^{2+} binding score, and indicating that the apo protein is likely to bind calcium.

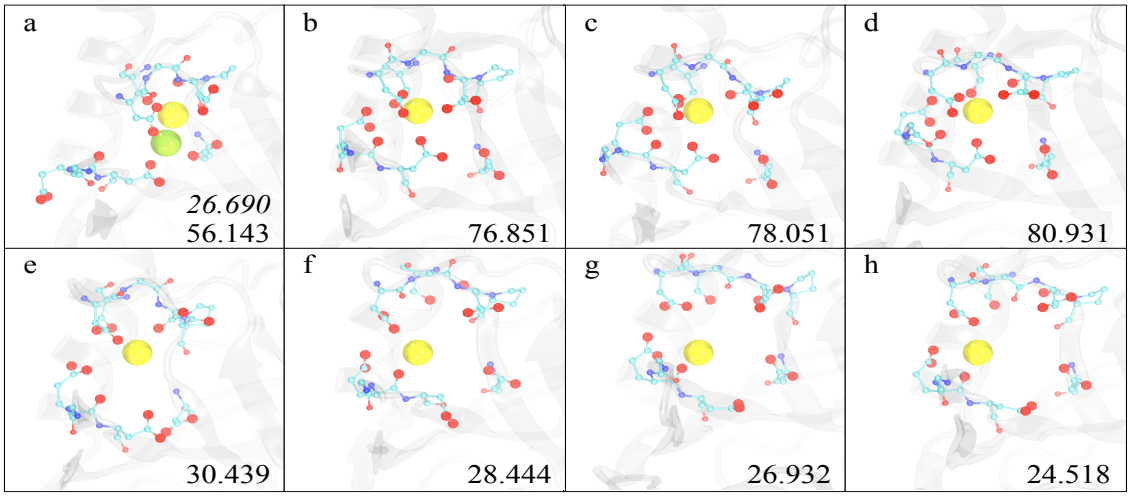


Figure 2. Neighbourhoods of the highest scoring points, obtained by scoring frames of MD simulations of PDB_ID:1I40 using Ca^{2+} binding FEATURE model, are depicted. In each panel, the score of the highest scoring grid point for the depicted frames is listed in plain print in the lower right corner. a) Original static PDB structure is shown with actual crystallized Ca^{2+} (in green) and with the highest scoring grid point (in yellow). The number appearing in the lower right corner in italics refers to the FEATURE score centered on the crystallized Ca^{2+} atom. b, c, d) Highest scoring frames in the simulation are depicted. e, f, g, h) Lowest scoring frames in the simulation are depicted. Gray - protein backbone, cyan - carbon, blue - nitrogen, red - oxygen, green - crystallized Ca^{2+} atom, yellow - highest scoring grid point. Large red spheres represent oxygens that coordinate Ca^{2+} binding; small red spheres represent other oxygens. Images were generated using Visual Molecular Dynamics (VMD) suite.

ENFIN: computers meet experiments

Ewan Birney

EMBL Outstation - Hinxton,
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton,, Cambridge, CB10 1SD, United Kingdom
birney@ebi.ac.uk

ENFIN is a European NoE combining bioinformatics groups and experimental groups working in systems biology. ENFIN has developed both new technologies for working with systems biology databases and new approaches in for predicting both discrete functions, network properties and system properties. I will illustrate ENFIN with a number of joint computational and experimental projects.

Scaling SIFTER to annotate larger, functionally diverse protein families

Barbara E Engelhardt[1], Michael I Jordan[1,2] John R Srouji[3] and Steven E Brenner[4*]

Departments of [1]Electrical Engineering and Computer Science, Statistics, Molecular and Cell Biology, and Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, USA. *Presenting author

email: Barbara E. Engelhardt bee@cs.berkeley.edu; Michael I Jordan - jordan@cs.berkeley.edu; John R Srouji -jrsrouji@compbio.berkeley.edu; Steven E. Brenner - brenner@compbio.berkeley.edu

Statistical Inference of Function Through Evolutionary Relationships[1,2] (SIFTER) uses a statistical graphical model to automate precise protein function annotation by applying principles from phylogenomics[3]. A previous version of SIFTER (version 1.1) performed well in predicting molecular function in a variety of protein families[2]. However, that version of SIFTER could only be applied to protein families with minimal functional diversity. We will present a new version of SIFTER (version 1.2) that is suitable for large and functionally diverse protein families because it includes a more general model of protein function evolution and a fast method for approximate calculation of posterior probabilities.[4]

Essential to the development of improved methods for protein function prediction is the availability of gold-standard test sets. We had previously developed the AMP/adenosine deaminase family as a gold standard using a score of experimental functional characterizations culled from an exhaustive search of the literature, in addition to the half-dozen seen previously in databases[1]. This family was excellent for assessing functional variation in a moderately large family with some limited functional diversity. For greater functional diversity we have established two more gold-standard protein families, the sulfotransferases (of which 48 have experimentally characterized molecular functions recorded in the GOA database) and the Nudix family for which we did a manual literature survey which revealed 97 proteins taking on 66 distinct experimentally-verified molecular functions.

We validated the new version of SIFTER on the previously studied the AMP/adenosine deaminases, and we investigated the performance on the two new protein families. SIFTER version 1.2 performed comparably to SIFTER version 1.1 as applied to the AMP/adenosine deaminase family of proteins, with 94% accuracy on an experimental data set. On the functionally diverse sulfotransferase protein family, SIFTER achieved 70% accuracy (where BLAST achieved 50%) on experimental data. The sulfotransferase family also showed that the new approximate computation of posterior probabilities works reliably across the full range of approximation granularities. On the exceptionally functionally diverse Nudix protein family, which was previously inaccessible to sifter because of the 66 possible molecular functions, SIFTER achieved 47% accuracy (where BLAST achieved 34%) on experimental data.

We also have demonstrated that a phylogenomic-based approach is effective on a genomic scale by applying SIFTER to 46 fully sequenced fungal genomes[5]. This new version of SIFTER, which has a more general model of function evolution and the ability to perform approximate computation of posterior probabilities, performs well on real protein data, as compared to sifter version 1.1 and other methods for protein function annotation. The approximate computation of posterior probabilities extends SIFTER's applicability to protein families with numerous, diverse molecular functions, such as the sulfotransferase and Nudix families.

4. REFERENCES

1. Engelhardt, B.E., Jordan, M.I., Muratore, K.E., and Brenner, S.E. 2005. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Computational Biology 1*:e32.
2. Engelhardt, B.E., Jordan, M.I., Brenner, S.E. 2006. A statistical graphical model for predicting protein molecular function. Proceedings of the 23rd International Conference on Machine Learning 038.1-038.8.
3. Eisen, J.A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research 8*:163–167.
4. Engelhardt, B.E., Jordan, M.I., Srouji, J.R., and Brenner, S.E. Scaling SIFTER to annotate larger, functionally diverse protein families. *in preparation.*
5. Engelhardt, B.E., Stajich, J. Jordan, M.I., and Brenner, S.E. Predicting protein molecular function for 46 fungal genomes using the SIFTER automated phylogenomic approach. *See abstract in this volume.*

Predicting functions of human proteins using patterns of native disorder

Anna Lobley[1], Christine A. Orengo[2], Mark B. Swindells[3], David T. Jones[12]*

1. Dept. of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
2. Dept. of Biochemistry, University College London, Gower Street, London WC1E 6BT, UK
3. Inpharmatica Ltd, 1 New Oxford Street, London, WC1A 1NU, UK

*To whom correspondence should be addressed:dtj@cs.ucl.ac.uk

INTRODUCTION

One of the challenges of the post genomic era is to predict the function of a protein given its amino acid sequence. Most automated function prediction methods rely upon identifying well annotated sequence and structural homologues to transfer annotations to uncharacterized proteins (see [1,2] for a comprehensive review). Sequence similarity based methods are relatively successful at annotating homologous proteins, however, they are not applicable to annotating orphan proteins or proteins whose relatives are not themselves functionally annotated. Currently, around 35% of proteins cannot be accurately annotated by homology-based transfer methods [3] highlighting the need for function prediction methods that are independent of sequence similarity.

Natively unstructured regions are a common feature of eukaryotic proteomes, particular those of higher organisms. Between 30 - 60% of proteins in eukaryotes are predicted to contain long stretches of disordered residues and not only have many of these regions been confirmed experimentally, but have also been found to be essential for protein function. In this study we directly address the potential contribution of protein disorder in predicting protein function using standard Gene Ontology (GO) categories. Initially we analysed the occurrence of protein disorder in the human proteome and identified ontology categories that are seen to be enriched in disordered proteins. To identify regions of disorder, we made use of DISOPRED2 [4] with a false positive rate of 2%. Pattern analysis of the distributions of lengths and positions of disordered regions in human sequences demonstrated that the functions of intrinsically disordered proteins are indeed both length and position dependent. These dependencies were then encoded in feature vectors to quantify the contribution of disorder in human protein function prediction using Support Vector Machine (SVM) classifiers.

An ensemble of SVM classifiers was trained to predict each GO term along similar lines to other automated function predictions methods such as ProtFun [5]. Each SVM utilizes different optimisation parameters and different feature weightings. Classification performance was evaluated over each of 5 different classifiers for each GO category. The prediction accuracies of 26 GO categories relating to signalling and molecular recognition were improved using the disorder features. The most notable improvements were observed for kinase, phosphorylation, growth factor and helicase categories. Furthermore we have generated predicted GO term assignments using these classifiers for a set of unannotated and orphan human proteins. We have benchmarked our method against a similar method for recognising protein function from sequence and report improved classification performance for all tested Molecular Function and Biological Process GO terms.

REFERENCES

1. Rost, B., Liu, J., Nair, R.,Wrzeszczynski, K.O. and Ofran, Y. 2003. Automatic prediction of protein function. Cell Mol.Life Sci. 60:2637-2650.

2. Friedberg, I. 2006. Automated protein function prediction--the genomic challenge. Brief.Bioinform. 7:225-242.

3. Ofran, Y., Punta, M., Schneider, R. and Rost, B. 2005. Beyond annotation transfer by homology novel protein-function prediction methods to assist drug discovery. Drug Discov.Today 10:1475-1482.

4. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., Jones, D. T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol., 337:635-645.

5. Jensen, L.J., Gupta, R., Staerfeldt, H.H. and Brunak, S. 2003. Prediction of human protein function according to Gene Ontology categories. Bioinformatics 19:635-642

A critical assessment of *M. musculus* gene function prediction using integrated genomic evidence

Lourdes Peña-Castillo1 , Murat Tasan2 , Chad L Myers3 , Hyunju Lee4 , Trupti Joshi5 , Chao Zhang5 , Yuanfang Guan3 , Michele Leone6 , Andrea Pagnani6 , Wan Kyu Kim7 , Chase Krumpelman8 , Weidong Tian2 , Guillaume Obozinski9 , Yanjun Qi10 , Sara Mostafavi11 , Guan Ning Lin5 , Gabriel Berriz2 , Frank Gibbons2 , Gert Lanckriet12 , Jian Qiu13 , Charles Grant13 , Zafer Barutcuoglu14 , David P Hill 15 , David Warde-Farley11 , Chris Grouios1 , Debajyoti Ray16 , Judith A Blake15 , Minghua Deng17 , Michael Jordan18 , William S. Noble19 , Quaid Morris1,11,20 , Judith Klein-Seetharaman21 , Ziv Bar-Joseph10 , Ting Chen4 , Fengzhu Sun4 , Olga G Troyanskaya3 , Edward M. Marcotte7 , Dong Xu5 , Timothy R. Hughes1,20 †, Frederick P. Roth2,22 †

1 Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, M5S 3E1 Canada 2 Dept. of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115 USA 3 Lewis-Sigler Institute for Integrative Genomics and Department of Molecular Biology, Princeton University, Princeton, NJ 08544 USA 4 Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA, 90089, USA 5 Digital Biology Laboratory, Computer Science Department and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO-65211, USA 6 ISI Foundation, Viale S. Severo 65, 10133 Torino, Italy 7 Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, 2500 Speedway, Austin, TX 78712 USA 8 Department of Electrical and Computer Engineering, Institute for Cellular and Molecular Biology, University of Texas at Austin, 2500 Speedway, Austin, TX 78712 USA 9 Department of Statistics, UC Berkeley, 367, Evans Hall #3680, Berkeley, CA 94720-3860, USA 10 School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA 11 Computer Science Department, University of Toronto, Toronto, ON, Canada 12 Dept. of Electrical and Computer Engineering, U.C. San Diego, Mailcode 0407, La Jolla, CA 92093-0407, USA 13 Genome Sciences, Box 355065, Foege Building, S220, 1705 NE Pacific St., Seattle, WA 98195-5065, USA 14 Department of Computer Sciences, Princeton University, Princeton, NJ 08544 USA 15 Bioinformatics and Computational Biology, The Jackson Laboratory, Bar Harbor, ME 04609 USA 16 Gatsby Computational Neuroscience Unit, London, UK. 17 School of Mathematical Sciences and Center for Theoretical Biology, Peking University, Beijing 100871, PRC 18 Department of Electrical Engineering and Computer Science, and Department of Statistics, University of California, Berkeley, EECS Department, 731 Soda Hall #1776, Berkeley, CA 94720-1776, USA 19 Department of Genome Sciences, and Department of Computer Science and Engineering, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195, USA 20 Banting and Best Department of Medical Research, University of Toronto, Toronto, ON, Canada 21 Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260 USA 22 Center for Cancer Systems Biology, Dana-Farber Cancer Institute, One Jimmy Fund Way, Boston, MA 02115 USA

† To whom correspondence should be addressed. Email: t.hughes@utoronto.ca (TRH) and fritz_roth@hms.harvard.edu (FPR).

1. INTRODUCTION

Several years after sequencing the human genome and the mouse genome, much remains to be discovered about the functions of most human and mouse genes. Computational prediction of gene function promises to help focus limited experimental resources on the most likely hypotheses. Several algorithms using diverse genomic data have been applied to this task, but these have been primarily tested on the unicellular yeast *S. cerevisiae*. In this study, a standardized collection of mouse functional genomic data was assembled, and nine bioinformatics teams used this dataset to independently train classifiers and generate predictions of function for 21,603 mouse genes. We identified strengths and weaknesses of current functional genomic datasets and compared the performance of function prediction algorithms. This analysis inferred functions for 76% of mouse genes, including five thousand currently uncharacterized genes, with an average precision value of 35% at a recall value of 20%.

The Evolution of Enzyme Mechanisms and Functional Diversity

Janet M. Thornton[a], Gemma L. Holliday[a], Irileria Nobeli[a], Rafael Najmanovich[a] , Daniel E. Almonacid [b], John B. O. Mitchell [b,] Angelo Favia [a], Structural Genomics Consortium[c]

[a] EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD,
[b] UK Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, UK,
[c] Structural Genomics Consortium, Rm 522 - Banting Building, University of Toronto, 100 College Street, Toronto, ONM5G 1L5 Canada

Enzyme activity is essential for almost all aspects of life. With completely sequenced genomes, the full complement of enzymes in an organism can be defined, and 3D structures have been determined for many enzyme families. Traditionally each enzyme has been studied individually, but as more enzymes are characterised it is now timely to revisit the molecular basis of catalysis, by comparing different enzymes and their mechanisms, and to consider how complex pathways and networks may have evolved.

To understand catalysis better we have developed MACiE (Mechanism, Annotation, and Classification in Enzymes), a database of the chemical mechanisms of enzymatic reactions. The MACiE dataset evolved from that published in the Catalytic Site Atlas (CSA) Porter *et al.* (2004)) and each entry is selected so that it fulfils the following criteria. There must be a 3-dimensional crystal structure of the enzyme deposited in the Protein Databank (PDB); the mechanism is relatively well understood mechanism; only one representative of each homologous protein family is included (H level of the CATH code – a hierarchical classification system of protein domain structures - unless there is a homologue with a significantly different chemical mechanism. Details of each proposed enzyme mechanism are stored in a MySQL database.

Research performed using MACiE to reveal which reaction steps are most common will be described.

ACKNOWLEDGEMENT
We would like to thank the EPSRC, BBSRC, IBM, Chilean Government and Cambridge Overseas Trust for funding and Unilever for supporting the Centre for Molecular Science Informatics and EMBL.

REFERENCES
Gemma L. Holliday, Daniel E. Almonacid, Gail J. Bartlett, Noel M. O'Boyle, James W. Torrance, Peter Murray-Rust, John B. O. Mitchell and Janet M. Thornton.
Journal: Nucleic Acids Res (Database Issue); D515-D520; 2007
Title: MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms.

Porter,C.T., Bartlett,G.J., Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucl. Acids. Res., 32, D129-D133.

The genome-wide annotation schema in the BioSapiens-ENCODE collaboration

Alfonso Valencia*, Martin Krallinger and Michael Tress
Centro Nacional de Investigaciones Oncológicas, Madrid, Spain
* To whom correspondence should be addressed: valencia@cnio.es

1. INTRODUCTION

A considerable number of platforms have been developed for the systematic annotation of protein function at the genome level based on the transference of annotations between related sequences (see Valencia, 2005).

During the annotation of the ENCODE (The ENCODE Project Consortium, 2004) protein dataset we entered in a new challenging area related with the annotation of potential protein splice variants.

The preliminary annotation of the 1097 proteins corresponding to well established genes in the 1% of the human genome (Tress *et al*., 2007) carried out by the Biosapiens Network (www.biosapiens.info) showed the capacity and limitations of current annotation methods.

The Biosapiens partners are now working on the organisation of a systematic annotation pipeline able to cope with the foreseeable need of annotating the potential set of alternative splicing forms that will be delivered by the ENCODE scale up project.

In a different, but related context our group,in collaboration with groups from NCBI, EBI, U. Rome and Mitre Corp., has recently organized the second Biocreative challenge for the assessment of text mining methods in biology. The analysis of the result in terms of the capacity of current methods to find functional information in text are highly relevant in the context of database annotation efforts. Furthermore the organisation of a public metaserver offering combined text mining results in the context of Biocreative will certainly enable the incorporation of text mining tools into automatic annotation pipelines.

2. REFERENCES

Valencia A. 2005. Automatic annotation of protein function. *Curr Opin Struct Biol.* 2005 Jun;15(3):267-74.

The ENCODE Project Consortium. 2004. *Science* 306:636-640.

Tress M.L., Martelli P.L., Frankish A., Reeves G.A., Wesselink J.J., Yeats C., Olason P.L., Albrecht M., Hegyi H., Giorgetti A., Raimondo D., Lagarde J., Laskowski R.A., Lopez G., Sadowski M.I., Watson J.D., Fariselli P., Rossi I., Nagy A., Kai W., Storling Z., Orsini M., Assenov Y., Blankenburg H., Huthmacher C., Ramirez F., Schlicker A., Denoeud F., Jones P., Kerrien S., Orchard S., Antonarakis S.E., Reymond A., Birney E., Brunak S., Casadio R., Guigo R., Harrow J., Hermjakob H., Jones D.T., Lengauer T., Orengo C.A., Patthy L., Thornton J.M., Tramontano A. and Valencia A. 2007. The implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad Sci U S A.* 104(13):495-500.

Functional annotations of protein 3D structures with the GENEFUN meta function assignment resource.

Lensink M.F.[1], Fischer D.[2], Yu C.S.[3], Dessailly B.H.[1], Saini H.K.[2], del Pozo A.[4], Tress M.[4], Valencia A.[4], Ekman D.[5], Björklund Å.[5], Elofsson A. [5] Reshef D.[6], Yahalom R.[6], Keasar C.[6], Raes J.[7], Bork P. [7], von Grotthuss [8] M.,Rychlewski L.[9], **Wodak S.J.[1,3*]**

[1]*SCMBB, Université Libre de Bruxelles, CP 263, Bd. du Triomphe, 1050 Brussels, Belgium.*
[2]*Center of Excellence in Bioinformatics and Dept. of Computer Science and Engineering, University at Buffalo, 901 Washington Street, Suite 300, Buffalo, NY 14203, USA*
[3] *Molecular structure and function program, Hospital for Sick Children, 555 University Ave., Toronto, Ontario M5G 1X8, Canada*
[4] *Structural Computational Biology Programme, Spanish Cancer Research Centre (CNIO), E-28029 Madrid, Spain*
[5] *Stockholm Bioinformatics Center, Stockholm University, SE-10691 Stockholm, Sweden*
[6] *Dept. of Computer Science, Ben-Gurion University, Beer-Sheva, Israel*
[7] *Structural and Computational Biology Unit, EMBL, Heidelberg, Germany*
[8] *Dept. of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA-02138, USA*
[9] *BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland*
Corresponding author: Shoshana J. Wodak, e-mail:Shoshana@sickkids.ca

1. INTRODUCTION

Structural genomics efforts are often producing 3D structures of proteins with unknown function. Since the protein 3D structure plays an important role in determining its function, the use of structural information should be helpful in predicting what this function might be. A number of methods, that exploit such information to generate functional annotations for proteins, have recently been proposed. But comparing their performance has so far been hampered by the fact they tend to use different functional descriptors and have been validated on different datasets.

2. METHODS

Partners of the GeneFun project have undertaken the development of a Web-resource (http://www.scmbb.ulb.ac.be/GeneFun/), which collates predictions made by different methods on a set of 259 protein 3D structures of unknown function, produced by structural genomics efforts in the US, Canada and Europe (Figure 1).

Several of these methods have been developed by GeneFun partners. Some infer the EC number or catalytic residues of a putative enzyme[1,2], others identify binding sites from the atomic coordinates, based on energetic criteria[3] or on evolutionary relationships[4-6]. Several methods assign functional categories from the Gene Ontology hierarchy[7]. In addition to the assignments by GeneFun partners, functional annotations are produced for the same set of structures by 3 publicly available Web servers, ProKnow[8], ProFunc[9] and Jafa[10]. The functional assignments produced for each structure by all these different methods are stored for browsing and further analysis.This analysis has so far focused on investigating the level of consistency between the predictions from different methods, with the ultimate goal of deriving a consolidated prediction approach. Another focus has been extending and validating EC number predictions by using more detailed information on active site residues from the CSA database[11], and by searching the PDB for highly similar active sites.

3.RESULTS

Results obtained so far indicate that the predictions obtained by different methods are sometimes inconsistent, but more often too general (as described by the GO hierarchy level, see Figure 2) to be useful. In a subset of cases, however, useful and consistent predictions are obtained. Our presentation will review and discuss these results.

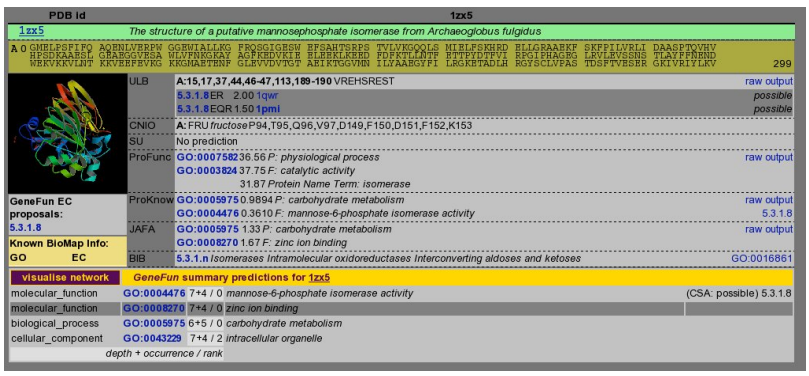


Figure 1: Screen shot of the GeneFun web resource summarizing the biological functions, predicted by different methods and research groups for 3D structures solved by structural genomics efforts, belonging to proteins of unknown function. The collated predictions (shown here for the PDB entry 1ZX5) are based on methods

developed by GeneFun partners (cited by the acronyms of their institutions: ULB, CNIO, SU, BIB), as well as from a number of other resources developed elsewhere (ProKnow, Jafa, ProFunc (see text)).

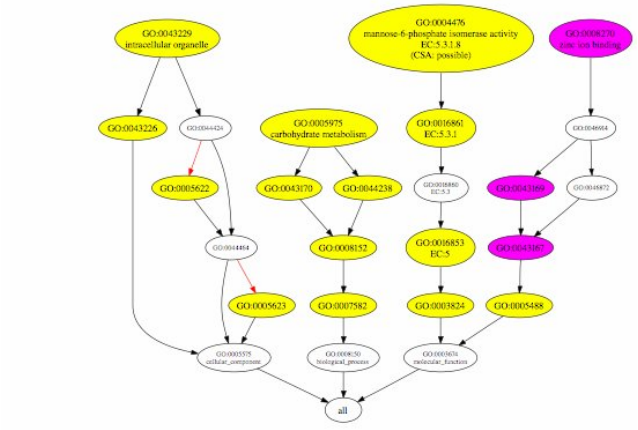


Figure 2: Predicted functions by GeneFun partners and other resources for the protein of hypothetical function highlighted in **Fig. 1** (PDB code 1ZX5). The predicted functions are displayed as Gene Ontology (GO) terms, mapped onto the GO hierarchy This hierarchy is displayed with least specific terms at the bottom and most specific at the top. These predictions fall into 2 main clusters based on the semantic similarity[12] of their GO terms. The two clusters are highlighted in Yellow and Magenta. GO terms ranked highly by any one method and therefore likely to be more reliable, are colored in a darker shade, but this is only barely visible in this picture.

4. ACKNOWLEDGEMENTS

We gratefully acknowledge support from European Commission FPVI for the project, In-Silico Prediction of Gene Function (GENEFUN), Contract Nº: LSHG-CT-2004-503567

5. LITERATURE REFERENCES

1. M. von Grotthuss, D. Plewczynski, K. Ginalski, L. Rychlewski, E.I. Shakhnovich, *PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics*, BMC Bioinformatics (2006), 7:53 [http://bioinfo.pl/PDB-UF/]

2. D. Reshef, R. Yahalom, N. Kalisman, Y. Gleyzer, C. Keasar, *Rare structural features as indicators of catalytic residues*, Structural Bioinformatics (2007), submitted
(this is the draft that Chen sent me, I was not able to find it in PubMed. The program will be part of the MESHI modeling suite: [http://www.cs.bgu.ac.il/~meshi])

3. B.H. Dessailly, M.F. Lensink, S.J. Wodak, *Relating destabilizing regions to known functional sites in proteins*, BMC Bioinformatics (2007), 8:141

4. C. von Mering, L.J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, P. Bork, *STRING 7-recent developments in the integration and prediction of protein interactions*, NAR (2007), 35:D358-362 [http://string.embl.de/]

5. F. Abascal, A. Valencia, *Automatic annotation of protein function based on family identification*, Proteins (2003), 53:683-692 [http://www.pdg.cnb.uam.es/funcut.html]

6. D. Ekman, Å. Björklund, J. Frey-Skött, A. Elofsson, *Multi-domain proteins in the three kingdoms of life - Orphan domains and other unassigned regions*, J.Mol.Biol. (2005), 348:231-243 [http://sbcweb.pdc.kth.se/cgi-bin/diaek/domfunction.cgi]

7. The Gene Ontology Consortium, *Gene Ontology: tool for the unification of biology*, Nature Genetics (2000), 25:25-29

8. D. Pal, D. Eisenberg, *Inference of protein function from protein structure*, Structure (2005), 13:121-130 [http://www.doe-mbi.ucla.edu/Services/ProKnow/]

9. R.A. Laskowski, J.D. Watson, J.M. Thornton, *ProFunc: a server for predicting protein function from 3D structure*, NAR (2005), 33:W89-93 [(http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/]

10. I. Friedberg, T. Harder, A. Godzik, *JAFA: a protein function annotation meta-server* NAR (2006), 34:W379-381 [http://jafa.burnham.org/]

11. C.T. Porter, G.J. Bartlett, J.M. Thornton, *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data*, NAR (2004), 32:D129-133 [http://www.ebi.ac.uk/thornton-srv/databases/CSA/]

12. Lord PW, Stevens RD, Brass A, Goble CA.
Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.
Bioinformatics. 2003 Jul 1;19(10):1275-83.

CONTRIBUTED TALKS

FunSite: a gold standard dataset of ligand-binding sites in protein structures

Dessailly BH* [1,2], Lensink MF [1], Orengo CA [2], Wodak SJ [1,3]
1 Service de Conformation des Macromolécules Biologiques, CP 263, Université Libre de Bruxelles (U.L.B), Bld. du Triomphe B-1050 Bruxelles, Belgium
2 Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK
3 Structural Biology and Biochemistry Program, Hospital for Sick Children, 555 University Avenue, Toronto Ontario M5G 1X8, Canada
*To whom correspondence should be addressed: benoit@scmbb.ulb.ac.be

1. INTRODUCTION

An increasing number of three-dimensional structures are determined for proteins for which nothing is known about the function. Among the different strategies used to predict function from structure, a common approach is to search for functional sites (1). But the absence of an appropriate gold standard dataset of functional sites in protein structures is a major limitation for the validation of functional site prediction methods. Whereas the Catalytic Site Atlas is now providing detailed descriptions of catalytic sites (2), reference gold standard datasets of binding sites are lacking.

Currently, binding site prediction methods are validated against datasets of known binding sites, which vary from study to study, making it difficult to compare the prediction performance of different methods.

In this work, we have identified several criteria which should be considered when building a gold standard dataset of binding sites in protein structures. Using these criteria we assemble a dataset of known ligand binding sites in proteins, and this dataset is compared to various datasets proposed so far.

2. FUNSITE: A GOLD STANDARD DATASET OF BINDING SITES

In this work, ligands are defined from the HETATM records of the PDB files. Our dataset therefore consists exclusively of binding sites for small ligands and polysaccharides.

Ligands in PDB structures are sometimes not biologically relevant and appear in the structure as a result of the experimental process of obtaining suitable crystals. In this work, we consider only ligands that make more than 70 inter-atomic contacts with the protein residues in order to avoid considering non-specific ligands. This inter-atomic contacts threshold was derived from a manual verification of more than 600 ligand-binding sites for their biological relevance. In other proposed datasets, the consideration of non-relevant ligands is avoided by simply excluding specific molecules from the analysis. This is not satisfactory since some of the excluded ligands can mimic biologically important molecules, and could hence provide useful information on relevant binding sites. In addition, the definition of biologically irrelevant molecules to exclude is not straightforward and may lead to erroneous exclusion or inclusion of compounds.

Binding site prediction methods should be applied to unbound structures, and validated against binding sites defined from the corresponding bound structures. This is important because proteins may undergo structural changes upon ligand binding. Our gold standard dataset consists of proteins with one high-resolution crystal structure of the unbound protein, and at least one bound structure. Protein residues are considered to be part of the binding site if they make contacts with the ligands, as defined by the software LPC (3).

Our dataset currently comprises 192 proteins, none of which share more than 25% sequence identity with any other. The binding sites are defined from the biological unit as defined by the PQS software (4). The dataset includes all the proteins in the December 22nd 2006 release of the PDB, which match our criteria and is therefore representative of the known data. When several bound structures are available for a given

protein, all are used to define the binding site. Overall, 484 bound structures are used to define the binding sites in our 192 proteins, and 86 proteins have at least 2 bound structures. On average, there are 27 residues making up binding sites per protein.

Our gold standard dataset does not require manual annotation, and can therefore be updated automatically as new structures become available.

3. CONCLUSION

FunSite, our gold standard dataset of ligand-binding sites in protein structures will be made publicly available. FunSite should complement well datasets describing other types of functional sites, such as the catalytic sites in the Catalytic Site Atlas (2), or the protein-protein interfaces in the CAPRI benchmark (5). Taken together, these datasets should improve our understanding of functional sites in proteins and help other researchers to derive and validate new functional site prediction methods.

4. ACKNOWLEDGMENTS

This work was supported in part by the GeneFun STREP project (contract N° LSHG-CT-2004-503567) funded by the European Commission, Framework Program VI.

5. REFERENCES

1. Jones S. and Thornton, J.M. 2004. Searching for functional sites in protein structures. *Curr Opin Chem Biol* 8:3-7.

2. Porter C.T., Bartlett G.J. and Thornton J.M. 2004. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32:D129-D133.

3. Sobolev V., Sorokine A, Prilusky J., Abola E.E. and Edelman M. 1999. Automated analysis of interatomic contacts in proteins. *Bioinformatics* 15:327-332.

4. Henrick K. and Thornton J.M. 1998. PQS: a protein quaternary structure file server. *Tr Bioch Sci* 23:358-361.

5. Mintseris J., Wiehe K., Pierce B., Anderson R., Chen R., Janin J. and Weng Z. 2005. Protein-protein docking benchmark 2.0: an update. *Proteins: Struct Funct Bioinf* 60:214-216.

Enriching functional knowledge in proteomics data using high-throughput PFP prediction

Troy Hawkins*, Meghana Chitale, Daisuke Kihara
Department of Biological Sciences, Purdue University, West Lafayette, IN, 47907, USA
Department of Computer Science, Purdue University
*To whom correspondence should be addressed: thawkins@purdue.edu

1. INTRODUCTION

We have previously designed and implemented a sequence-based Protein Function Prediction system, PFP (http://dragon.bio.purdue.edu/pfp), which annotates a query sequence using Gene Ontology (GO) terms in each of the three categories (1). PFP is aimed at high-coverage, lower resolution function prediction for large genomic and proteomic datasets, where a limiting factor in their interpretation is a lack of any functional annotation for nearly half of the included sequences. The PFP algorithm interprets PSI-BLAST search results by scoring GO terms from individual sequence hits by alignment strength (E-value) and frequency of term occurrence among all hits. Additionally, we incorporate data mining methods to predict additional GO terms based on co-occurrence association in UniProt sequences. PFP outputs a list of GO terms predicted for a query sequence. For each term, we calculate a p-value significance score and an estimated accuracy for the prediction. To evaluate the effectiveness of annotating protein sequences using PFP, we constructed a data set of target sequences consisting of sequences obtained from GOA associated to at least one GO term for eleven diverse proteomes. The evaluation consisted of stripping the annotations from these sequences and quantifying the recovery of those annotations by PFP.

A unique characteristic of PFP is the mining of functional information from divergent sequence hits retrieved by PSI-BLAST, i.e. those with E-values well above commonly accepted thresholds for significance. In order to assess the importance of utilizing the information found in these sequences, we evaluated sequence coverage while ignoring significant sequence hits with E-values below 8 cutoff values. When the complete PSI-BLAST results were used (E-value cutoff of 0.0, disregarding self-hits), PFP recovered biological process terms correctly for 72% of the benchmark sequences. As expected, this coverage drops as the most significant hits are ignored. Interestingly, however, even when only slightly similar sequences were used, we could still predict correct terms for one in five query sequences. We use simple transfer of GO terms from PSI-BLAST hits as a baseline for performance of PFP. We compared PFP predictions of GO biological process terms to PSI-BLAST at all E-value cutoffs. When only sequence hits of E-value ≥ 1 were used for making predictions, the coverage of benchmark sequences by PFP more than doubles that of PSI-BLAST annotation alone. PFP has a distinct advantage in being able to predict more general GO terms when a specific biochemical activity or biological process cannot be predicted from similar sequences. This is apparent when we analyze the average depth of correct predictions made by PFP at all E-value cutoffs.

We assigned significance scores to each of the predicted GO terms output by PFP and related those scores to an expected accuracy for blind predictions. For each term we determined the distribution of raw scores, which was used to assign a term-specific p-value to each prediction. This value represents the significance of a particular score relative to its distribution; however, for any given term the relationship between statistical significance and accuracy is unique. We constructed standard curves relating p-value significance and specificity for each GO term. The overall accuracy of PFP for GO molecular function annotations across the benchmark set is represented well in a precision-recall plot (Figure 1). It should be noted that this plot relates specificity and sensitivity at the annotation level, i.e. sensitivity represents the percentage of all known annotations recovered at each p-value threshold, not just the percentage of sequences for which we could predict a single correct term (sequence coverage).

2. FUNCTIONAL ENRICHMENT OF PROTEOMES AND PROTEIN INTERACTION NETWORKS

With established significance scores and a method for relating p-value to an estimated accuracy, we were able to assign predictions to unannotated proteins in fifteen genomes and assess the ability of PFP to enrich the functional knowledge of these proteomes at any given confidence level. For each proteome, we counted

the number of unknown proteins for which we could make a molecular function prediction with an estimated accuracy of greater than 90%. In each of the fifteen organisms we looked at, more than half of the previously unknown proteins could be assigned a GO molecular function term at the highest confidence level, and nearly 100% of these proteins could be assigned a term with an estimated accuracy of 40% or higher (Figure 2). We applied PFP to unknown proteins in protein interaction networks for *P. falciparum* (malaria plasmodium), *D. melanogaster* (fruit fly), and *C. elegans* (roundworm). Interactions here can be divided into three categories: (i) fully enriched, i.e. those where both proteins have some known or electronically assigned function, (ii) partially enriched, i.e. those where only one of the two proteins has some known function, and (iii) those where neither of the proteins has some known function. For the *P. falciparum* network (2), we could increase the number of fully enriched interactions by more four-fold, to over 93% of the total interactions, using GO biological process predictions with an estimated accuracy of over 90%.

3. FIGURES

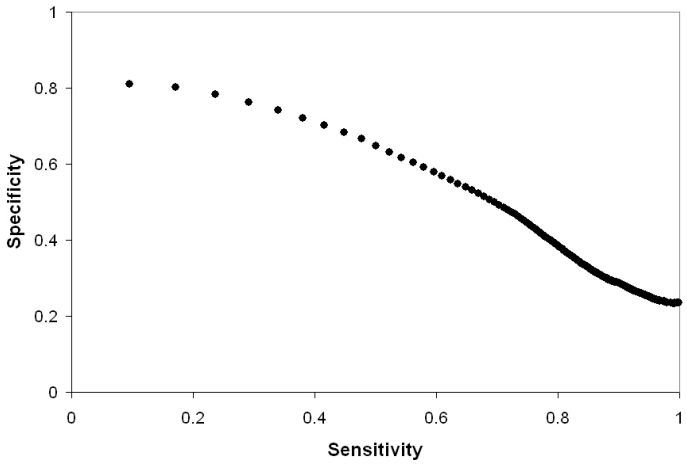


Figure 1. Precision-recall of predictions by p-value significance score. Precision (Y-axis, percentage of correctly predicted annotations) and Recall (X-axis, percentage of known annotations predicted correctly) are shown at thresholds of p-value significance stepped by 0.005. Predictions made with p-value of 0.005 or better are represented by the data point at the top left; those made with p-value of 1.000 or better (all) are represented by the data point at the bottom right.

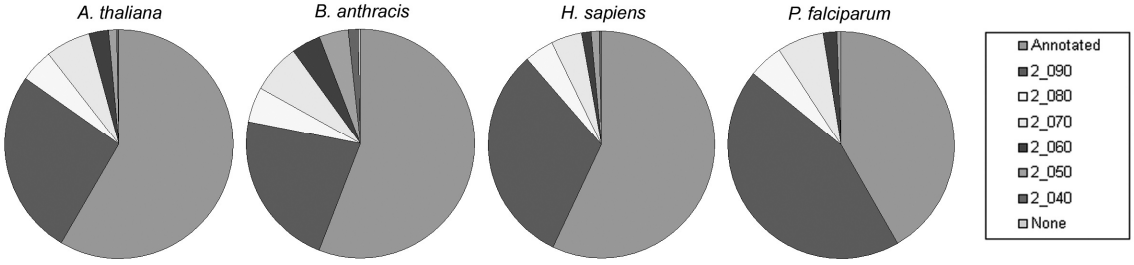


Figure 2. Functional enrichment of proteins from four organisms using predictions by PFP, by expected accuracy. Large light-gray pie pieces (Annotated) represent previously annotated proteins, large dark-gray pie pieces (2_090) represent proteins for which a GO molecular function term can be predicted with ≥ 90% expected accuracy.

4. REFERENCES

1. Hawkins, T., Luban, S. and Kihara, D. 2006. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Science* 15(6): 1550-1556.

2. LaCount, D.J. *et al*. 2005. A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* 438(7064): 103-107.

Protein similarity profile based functional annotation algorithm

Inkyung Jung[1], Jungsul Lee[1], Chulhee Choi[1], Dongsup Kim[1]*,
[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, 305-701, Korea.
*To whom correspondence should be addressed: kds@kaist.ac.kr

1. INTRODUCTION

The explosion of raw sequence data requires efficient computational methods for functional annotation. The functional relationship between proteins has been studied by probing evolutionary relationship between proteins (1) and using specific structural or functional information. Many effective approaches have been developed, yet still there is an ample room for improvement. In this work, we propose a simple but, powerful method for functional annotation through combining two novel algorithms; one is to reduce "noises" that may originate from incorrect alignments or improper scoring system of certain functional annotation methods and, the other is to develop a new way to measure similarity between query and template profiles.

SWISSPROT-EC mapping database (2) is used to select query proteins (2571) and template proteins (5907). For a given template set, we first construct the protein similarity score profiles (PSSPs) for all templates. In these profiles, each entry is a similarity score between protein *i* and *j*, $\exp(-\min(e_{ij},e_{ji})/95)$, where e_{ij} is an E-value calculated by PSI-BLAST (3). Second, we rank all templates according to their E-values for a query sequence. Next, we create the "information profile" by summing the PSSPs below the E-value threshold, 3, and the "background profile" by the same way using proteins over the E-value threshold, 700. Information profile is considered to be a "signal" that well represents the similarity between a query and templates, whereas background profile is believed to contain information on "noises". We build a query profile by subtracting the background profile from the information profile with different weight, 0.7 for the information profile and 0.3 for the background profile. As a result, noise-reduced query profile is constructed, which gives new similarity scores for templates. We also build a template profile for all templates as the same way for a query profile. In our method, PSSPs and template profiles are prerequisite, which need to be calculated only once. Figure 1 depicts the procedure of building a noise reduced query profile.

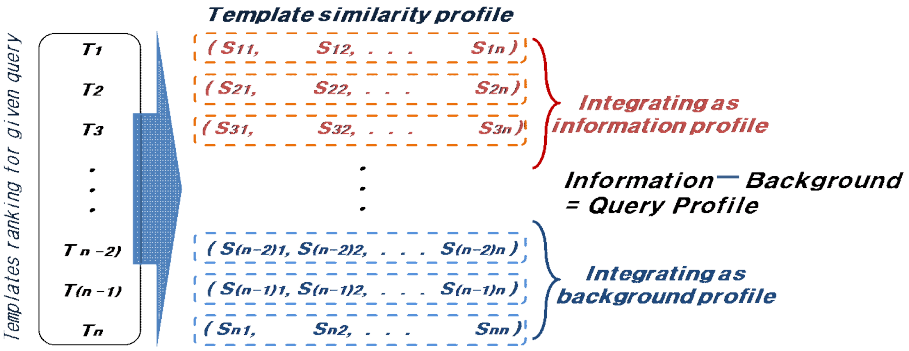


Fig1. Illustration of building query profile scheme. T*i* represents template *i* and, S*ij* means the similarity score between templates *i* and *j*. Subtracting background profile from information profile effectively reduces "noises".

We then measure the similarity between a query profile and template profiles, which allows new ranking system for templates for a given query. We develop a profile similarity measurement by considering the tendency of similarity scores. For a pair of a query profile (Q) and a template profile (T*i*), the inner product of Q and T*i* gives profile similarity. The results of inner product are normalized to remove the relative effects among template profiles. A similarity measurement between Q and T*i* is defined by

$$Similarity(Q,T_i) = \frac{Q \bullet T_i - \sum_{k \in T} Q \bullet T_k / \|T\|}{SEDEV(Q)} \quad \textbf{(1)}$$

,

where, T is a template set, Q • T*i* is inner product of a query profile and a template profile *i*, and STDEV(Q) indicates standard deviation of inner product of a query profile and all template profiles.

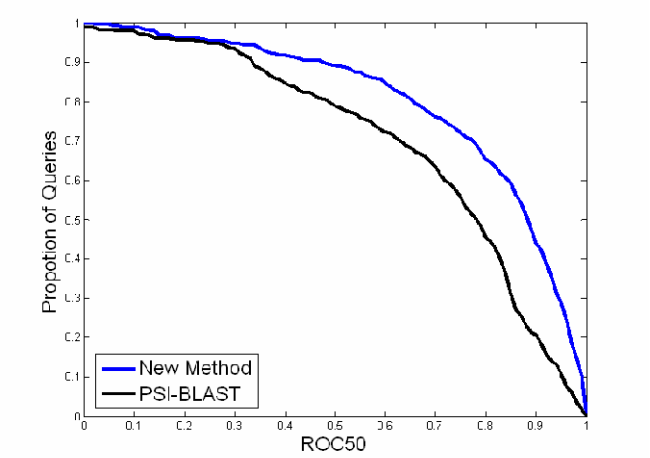


Fig2. Performance variation with various methods at EC fourth level. The x-axis represents ROC50 scores and the y-axis represents the proportion of queries which have better performance than the corresponding ROC50 score. The mean ROC50 score for two methods are 0.80 (New method, blue line) and 0.70 (PSI-BLAST, black line), respectively.

Performance of our method is assessed by comparing the fourth level of the EC hierarchy assigned by our method with those recoded in the SWISS-PROT entry. ROC50 scores are used for performance measure. Comparing the result of our method to that of PSI-BLAST, ours outperforms PSI-BLAST. We detects 44% of proteins in the same functional category at EC fourth level with >0.9 ROC50 score, while PSI-BLAST detect only 20% of them (Fig 2). For example, PSI-BLAST recognizes Ca2+-transporting ATPase proteins (EC 3.6.3.8) as highly related to glucokinase proteins (EC 2.7.1.2). Due to the tendency of Ca2+-transporting ATPase proteins to be assigned to low E-value as a result of improper scoring system of PSI-BLAST, Ca2+-transporting ATPase appears to be functionally related with many unrelated proteins. However, our method reduces false positives (e.g., Ca2+-transporting ATPase proteins) with background information, revealing more functionally related proteins.

This finding suggests that the performance of function prediction can be improved by considering the general context of protein similarity profiles, reducing “noises” using background profile and applying new scoring scheme to assess similarity between query and template profiles. By identifying the relationship between sequences and functions, our method aids to reveal the principle of protein function.

4. REFERENCES

1. Engelhardt B. E., Jordan M. I., Muratore K. E. and Brenner S. E. 2005. Protein Molecular Function Prediction by Bayesian Phylogenomics, PLoS Comput Biol 1(5)

2. Martin A. C. R. 2004. PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt, BIOINFORMATICS, 20, 986-988.

3. Atschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res, 25, 3389-3402

Functional Prediction of Abandoned Short Sequences: Toxin-like Peptides from Insect and Mammalian Genomes

Michal Linial*, Noam Kaplan and Noa Morpurgo
Dept of Biological Chemistry, The Hebrew University of Jerusalem, 91904, Israel
*To whom correspondence should be addressed: michall@cc.huji.ac.il

1. INTRODUCTION

Numerous high throughput (HTP) methodologies have been developed in recent years. The application of such technologies drastically changed the scale and the pace of discovery in biological and medical research. As electronic inference for genome annotation became dominated, hundreds of sequences may have been overlooked and obviously are ignored by most current proteomics technologies. We will discuss methods that support data analysis from mixed sources for improving the confidence in functional inference [1]. The application of these methods to the experimentalists will be discussed [2].

Herein we focus one group of abandoned short sequences many of them belong to animal toxins. Toxins are extremely varied in their structure and biochemical function. The function of these short peptides as toxins seems unclassifiable by sequence-based methods. In spite of this, many toxins share a structural stability property, which is likely required for traveling through the hostile blood stream of the venom recipient. We restrict our predictions to animal peptide toxins (APT) that are short proteins appearing in animal venom and that are aimed at inflicting harm to the organism on which the venom acts. APTs are extremely varied in terms of function and include ion channel inhibitors (ICIs), phospholipases, protease inhibitors, disintegrins, defensins and other biological groups. Even specific groups of ICIs, which inhibit the same target channels, often vary in sequence and structural fold.

Most state-of-the-art functional classification methods use position specific information (e.g. evolutionary conserved positions) in order to find sequence motifs that are common to functional groups. Due to the large variation of APTs in sequence and structure, this commonly used approach is unsuitable in the case of APTs. Our classifier uses 545 general sequence-derived features which we had speculated to possibly be related to APT structural stability. The features were constructed so that they would reflect the frequency, distribution, packing and crude localization of cysteines within the sequence. However, the features were not restricted to cysteine-related features and were applied to all 20 amino-acids.

2. RESULTS

We present the construction of a toxin classifier by extracting sequence-derived features that are guided by the notion of structure stability. We focus on short peptides that share properties with most ion channel inhibitors. The method is able to recognize high-level functionality of toxin-like functions. We implemented the method into a ToxClass, a tool that analyses a set of sequences and provides a ranked score for the prediction of each sequence to act as a toxin-like protein.

Several groups of proteins were identified with a positive score by our classifier. In addition to toxins, it is apparent that various antibacterial groups are over-represented, specifically β-defensins. Venomous animals are subject to an evolutionary pressure to maintain their venom effectiveness. In cone snails, snakes and spiders, extensive duplications and mutations events in venom proteins have been observed, resulting in a rich mixture of peptide variants that cover a broad range of specificities towards their preys. Similarly, antibacterial peptides must cope with the fast evolution of the microbial world. We suggest these findings to indicate a close evolutionary relationship between APTs and antibacterial proteins, including possible evolutionary recruitment of antibacterial proteins to act as venom toxins or vice versa.

One over-represented biological group is that of the metallothioneins. Metallothioneins are ubiquitous cysteine-rich proteins that have been suggested to possess a variety of functions including zinc homeostasis and antioxidative effects. The full range of functions of these proteins remains unknown. However, antibacterial activity of a metallothionein protein expressed in housefly larvae has been reported recently. In summary, the classifier is apparently able to correctly produce a non-trivial characterization of APT and APT-like proteins.

When applied the classifier to 10157 predicted protein sequences from the recently sequenced honey bee (*Apis*

Mellifera) genome, several novel APT-like sequences were identified, including homologs from non-venomous species. This prompted us to search a non-venomous organism for APT-like proteins. In a recent paper, the mouse FANTOM consortium produced a large set of full-length cDNA sequences, producing an extensive representation of the mouse transcriptome. Amongst these were 5154 putative novel proteins to which we have applied our classifier, resulting in an additional novel family of mammalian APT-like proteins.

In the bee, one of the proteins that are predicted positive shows strong evidence of similarity to bug toxins and cone snail toxins, all of which function as voltage-gated Ca^{2+} channel inhibitors. We called this proteins OCLP1 (for omega conotoxin like protein, Figure 1). Strikingly, this protein seems to be expressed outside of the venom gland. A toxicity assay in which the protein was injected to fish induced a strong yet reversible paralytic effect. We suggest that the protein may function as an endogenous modulator of voltage-gated Ca^{2+} channels [3]. Furthermore, Raalin, an overlooked toxin-like peptide from bee is overexpressed in the brain and its sequence is conserved in numerous insects.

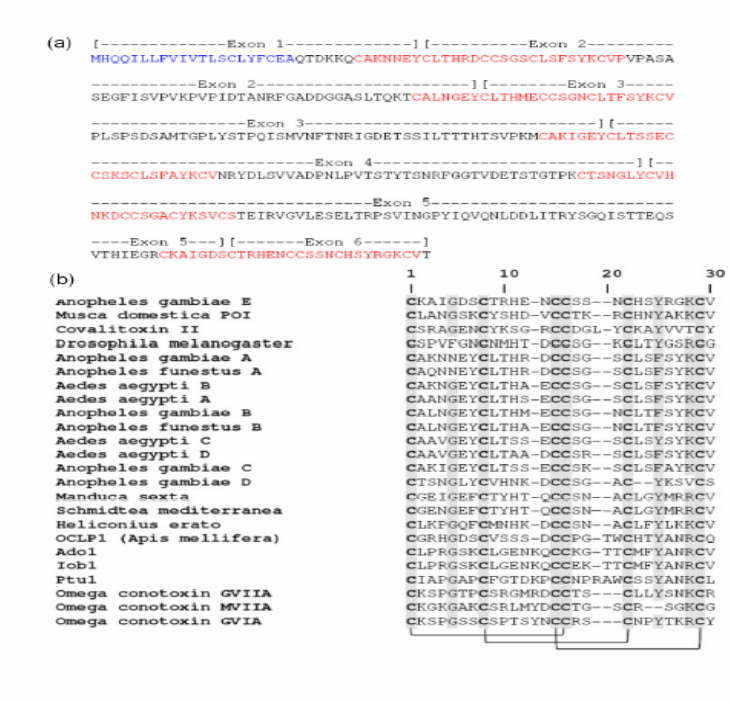


Figure 1. OCLP1 Homologs (a) Amino acid sequence of the *Anopheles gambiae* OCLP1 homolog. Blue amino acids represent the putative location of the signal peptide (predicted by SignalP). Red amino acids represent the locations of the OCL repeats. Note that the exons are positioned similarly relatively to the OCL repeats, with each of the exons ending before the second cysteine of an OCL repeat (see figure 2a).
(b) Multiple sequence alignment of OCL proteins. A-E indicates repeats within the OCLP1 protein homologs. Highly conserved positions are highlighted. Cysteines appear in bold. Disulfide connectivity is shown beneath the alignment. OCLP1 homologs are noted in species names only. A-E indicates OCL repeats. Only the OCL region is shown. Note the YANRC sequence which is shared only by OCLP1, Ado1, Ptu1 and Iob1.

In applying the classifier on the mouse transcriptome, we have identified a novel mammalian cluster of toxin-like proteins that are expressed in the testis and in brain. We suggest that these proteins might be involved in regulation of nicotinic acetylcholine receptors that affect the acrosome reaction and sperm motility. Finally, we highlight a possible evolutionary link between venom toxins and antibacterial proteins. We expect our methodology to enhance the discovery of additional novel protein families. The role of toxin-like function in non venomous context and the experimental challenges to detect these peptides will be discussed.

3. REFERENCES

1. Kaplan N, Vaaknin A, Linial M. (2003) PANDORA: keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res. 31:5617-5626.*
2. Sasson O, Kaplan N, Linial M. (2006) Functional annotation prediction: All for one and one for all. *Protein Sci. 15:1557-1562.*
3. Kaplan, N, Morpurgo, N, Linial, M. (2007) Novel short toxin-like proteins expressed in insects and mammals - Computational based discovery. *J. Mol. Biology, Mar 15 (in press)*

Conserved hydrogen bond patterns reveal structural and functional motifs in transmembrane protein regions

Annalisa Marsico, Andreas Henschel*, Gihan Dawelbait, Christof Winter, Anne Tuukkanen and Michael Schroeder
Biotechnological Center, TU Dresden, Tatzberg 47-51, 01307 Dresden, Germany
*To whom correspondence should be addressed: ah@biotec.tu-dresden.de

1. INTRODUCTION

Detailed analysis of structural features in transmembrane proteins is widely needed for enhancing understanding of many basic phenomena underlying cellular functions. Topology prediction methods alone cannot model the heterogeneous structural complexity of membrane proteins in an exhaustive way. Proteins are known to be rich in small 3D structural motifs important for protein folding and stability as well as for function (in active sites). Among them, Schellmann motifs, beta bulges, alpha turns, nest motifs and many others are very well characterized in terms of backbone torsion angles and hydrogen bond patterns (5-7). In addition, membrane proteins are rich in reentrant regions, interfacial helices, irregular structures at the water-membrane interface, structured loops (8).

Even if the structural role of several small 3D motifs has been widely recognized, their functional role is not always known. In some cases structural motifs have been found to be functionally very important, for example: nest motifs with their anion binding NH groups (5) (often occurring as part of small hydrogen-bonded motifs) are prominent in functional regions (like in the P-loops of nucleotide triphosphate-binding proteins); structured loops in membrane receptors bind activation ligands ; reentrant regions have functional roles as selectivity filters in channels (8).
Exhaustive studies on small structural motifs in globular proteins have been carried out, revealing interesting correlations between hydrogen bonds patterns, structural features and amino acid composition of these motifs (1, 2).

In this work we focus on alpha-helical transmembrane proteins and we systematically discover 3D motifs on the basis of common hydrogen bonds and characterize them in terms of sequence patterns, structural properties and functional relevance
We retrieve 125 non-redundant high-resolution alpha-helical membrane protein chains from the PDBTM database (3) and we generate residue fragments of different lengths ranging from 3 to 14 amino acids. We assign to each fragment: information relative to its location respect to the lipid bilayer, using the annotation stored in the PDBTM database and patterns of intra-segment hydrogen bonds (further distinguished on the basis of the atom type) retrieved by means of the Chimera algorithm. We perform hierarchical clustering for fragments of the same size and belonging to the same region on structural basis, using a similarity measure proportional to the absolute number of hydrogen bonds that two fragments have in common.
From the clustering procedure we generate a library of functionally annotated motifs that show specific hydrogen bond patterns and sequence conservation.
The number and distribution of representative clusters strongly depends from the considered fragment size and the cellular region. For membrane embedded regions 80% of the items belong to the dominant class of alpha-helices. The interface, cytoplasmic and extracellular region show more structural variability: besides the alpha-helices class we observe other classes, smaller in size, corresponding to irregular short helices, structured loops, Schellmann motifs, beta-turns, beta-bulges, alpha turn and also new uncharacterized structural motifs.

We relate the motifs defined from each subcluster to functional annotation in three different ways: 1. we look for GO annotation of the proteins the fragments belong to and we show that fragments belonging to the same class, if functionally related, they share the same GO annotation; 2. we automatically retrieve Swissprot functional annotation for the fragments, in order to discriminate motifs that play a pure structural role from those ones clearly related to

a given function; 3. we use the SCOPPI database (4), that classifies all domain-domain interactions observed in PDB structure files, in order to discriminate motifs at protein-protein interfaces.
Our method allows to discover structural motifs that enrich the topological description of membrane proteins and sheds light on mechanisms of membrane proteins through the functional characterization of 3D motifs. We perform structural similarity based annotation transfer within a single cluster and we will annotate membrane protein sequences in different genomes by using an opportune codification of the 3D motifs at sequence level.

2. FIGURES

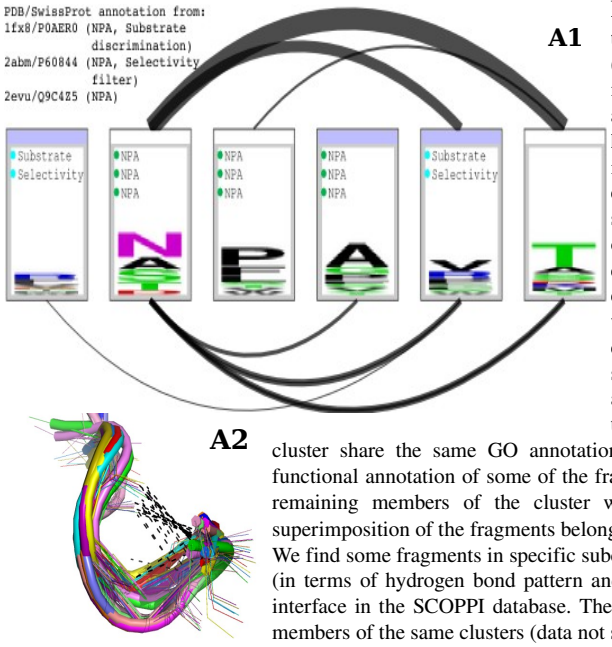


Fig.1: Examples. A1: significant structural motif derived from the clustering of 15 fragments of size 6 in the 'L' region (membrane embedded regions that do not entirely cross the membrane). Each rectangular box in the picture corresponds to a given residue's position . The motif shows a highly conserved hydrogen-bond pattern between residue's main chain atoms i+1, i+4 and i+1,i+5 (thicker arcs in the picture correspond to highly conserved hydrogen bonds). For all the cluster's items this structural motif is associated to a reentrant region in proximity of the selectivity filter of water channels and some cation channels. Boxes' tops are colored on the basis of the conservation score of a residue: the white ones correspond to well conserved amino acids and the blue ones to the weakly conserved ones. It is easy to recognize in this cluster the NPA sequence motif, signature of the aquaporin-like protein family and a conserved small amino acid (T or G) in position i+5, typical of a reentrant-like region. All the members inside the cluster share the same GO annotation relative to 'cation transport' or 'ion channel activity'. The functional annotation of some of the fragments in the subcluster suggests transfer of annotation to the remaining members of the cluster where no Swissprot annotation is available. **A2)** Structural superimposition of the fragments belonging to this class with the conserved hydrogen bonds patterns. We find some fragments in specific subclusters corresponding to very well conserved structural regions (in terms of hydrogen bond pattern and average RMSD) that are annotated as part of an interaction interface in the SCOPPI database. The annotation of interface region can be transferred to the other members of the same clusters (data not shown, available on request).

3. REFERENCES

1. http://www.ebi.ac.uk/msd-srv/msdmotif/

2. Han, K.F., Bystroff, C. and Baker, D. 1997. Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci .* 6, 1587-90.

3. Tusnady, G.E., Dosztanyi Z. and Simon, I. 2004. Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics.* 20, 17:2964-2972.

4. Winter, C., Henschel, A., Kim, W.K. and Schroeder, M. 2006. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acid Research.* 34, Database issue: D310-D314.

5. Watson, J.D. and Milner-White, J. 2002. A novel Main-chain Anion-binding Site in Proteins: The Nest. A Particular Combination of Φ, Ψ Values in Successive Residues Give Rise to Anion-binding Sites that Occur Commonly And Are Found Often at Functionally Important Regions. *J Mol Biol.* 315, 171-182.

6. Duddy, W.J., Nissink, J.W.M., Allen, F.H. and Milner-White, E.J. 2004. Mimicry by asx- and ST-turns of the four main types of β-turn in proteins. *Protein Sci.* 13:3051-3055.

7. Aurora, R. and Rose, G.D. 1998. Helix capping. *Protein Sci.* 7:21-38.

8. Viklund, H., Granseth, E. and Elofsson, A. 2006. Structural Classification and Prediction of Reentrant Regions in α-Helical Transmembrane Proteins: Application to Complete Genomes. *J Mol Biol.* 361, 591-603.

OntoDas - integrating DAS with ontology-based queries

Kieran O'Neill[*1], Anita Schwegmann[1,2], Rafael Jimenez[1,3], Dan Jacobson[1] and Alexander Garcia[1,4]
[1] Central Node, National Bioinformatics Network, Cape Town, 7405, South Africa
[2] Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, South Africa
[3] Proteomics Services Group, European Bioinformatics Institute, Hinxton, United Kingdom
[4] Centro Internacional de Agricultura Tropical, Cali, Colombia
*To whom correspondence should be addressed: kieran@nbn.ac.za

1. INTRODUCTION

We present OntoDas, an extension to the Dasty2 DAS client (1) which visually enables the construction of ontology-based queries for retrieving sets of related proteins. OntoDas is based on AJAX (asynchronous JavaScript and XML) and makes extensive use of web services, including Distributed Annotation System (DAS) (2), Ontology Lookup Service (OLS) (3) and a custom-built web service for executing the queries. By integrating multiple web services and making use of AJAX technology, we have created a tool that provides a unified view for creating dynamic, visual, ontology-based queries that is easy to use and install in other web-based systems. OntoDas facilitates the discovery of sets of proteins annotated with specific sets of Gene Ontology (GO) terms. The tool makes use of query previews to enable users to rapidly find results and to explore the query space.

Ontologies such as Gene Ontology (GO) (4) provide a means of integrating biological information from diverse sources. GO has been used to annotate gene products in numerous model organisms, and thus can and has been used as a platform for cross-database queries. A useful type of query which can be performed against GO and other ontologies is to find sets of gene products sharing annotations. An example might be: "*Retrieve gene products that participate in blood coagulation and are located in the extracellular space and have protease inhibitor activity.*" These queries can be executed using scripts, such as the Perl API to the GO MySQL database. However, both the task of constructing queries in a scripting language and the task of making sense of the results place high cognitive load on the user.

Tools such as GViewer (5) enable the execution of these kinds of queries. However, the interface for constructing queries is form-based, and requires the ontology terms to be known in advance by the user. This presents two problems: Firstly, there is no guarantee that the text entered will actually match an ontology term, and secondly that, even if it does match, the combination of terms will return any hits (6). This can be overcome by providing a query building interface, in which query previews (summarized previews of query results) are provided (6). In a query preview interface, options for modifying the query are provided, but constrained to valid terms in the ontology, and to those terms which, when added to the query, return some results.

Another feature which could aid users when constructing these kinds of queries is to enable them to build the query by using genes or gene products. For example, if the terms used to annotate a single protein are displayed, users may select combinations of those terms to specify queries to find related proteins. A tool such as Dasty2 for viewing the details for a single protein could be extended to be an entry point for an ontology based query system, and be used to view details on results returned.

OntoDas is a tool for visually constructing ontology based queries using GO. It makes use of Dasty2 as an entry point to query construction, and as a viewer for details on individual results. OntoDas employs information visualization techniques to assist biologists in creating and exploring queries to find sets of related gene products.

2. DESIGN OF ONTODAS

OntoDas was developed as an extension to Dasty2, an AJAX-based DAS client for viewing the sequence annotations of a single protein. As such, it makes extensive use of the existing paradigm and visual style of Dasty2, and employs Dasty2 as an entry and exit point to queries. In order to enable the lookup of ontology terms annotated to the protein being viewed, and to construct a query from a combination of these terms a panel was added to Dasty2. This takes the user to the main OntoDas view, which provides the user with the ability to retrieve query results, and to modify the query in useful ways that will produce nonempty result sets.

OntoDas aims to provide as much useful information as possible, without overwhelming the user. Queries are framed in natural language so as to make them easier to understand. This representation uses phrases based upon the formal relations laid out by Smith et al (7). The phrase “participate in” is used to refer to terms from the biological process ontology, to express the relation “has_participant”. The phrase “are located in” refers to terms from the cellular component ontology, in order to express the relation “has_location”. For terms from the molecular function ontology, since no relation had been agreed on, the phrase “have” is used, expressing a simple, non-specific property relation. Examples are shown in the introduction and in the screenshots in figure 1.

Additionally, “information scent” has been provided to guide users in choosing terms: Information scent is defined as “the (imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues” (8), and has been shown to be a highly important factor when finding terms in an ontology (9). OntoDas provides information scent by displaying complete term definitions when the user hovers the cursor over a term, and by showing the size of potential result sets. Full term definitions are used because terms themselves are often ambiguous, whereas their natural language definitions are more likely to ensure terms' appropriate interpretation (10). Finally, grouping and sorting of terms is provided to reduce the cognitive load on users in finding specific terms of interest (6).

OntoDas uses other web services in addition to DAS, specifically a custom-made web service acting as a query execution engine, and Ontology Lookup Service (OLS) (3), a powerful support vector machine-based text search tool for finding ontology terms. OLS provides the lexically similar neighbours which are displayed when a user is modifying a query term. The custom web service currently works with the GO relational database, using a small Python script front-end. It is designed to be modular, and potentially to use the DGB graph database developed by the NBN (11). By using multiple web services, powerful functionality can be provided from multiple sources in a single view. Similarly, by using JavaScript OntoDas enables visual manipulations without round trips to the web server, thereby improving the responsivity of the interface.

4. REFERENCES

1. Jimenez, R.C.; Quinn, A.F.; Labarga, A; O'Neill, K.; Garcia, A. and Hermjakob, H., 2007, Dasty2, a web client for visualizing protein sequence features (poster), *AFP-Biosapiens SIG, ISMB 2007*, (accepted)

2. Dowell, R.; Jokerst, R.M.; Day, A.; Eddy, S.R. and Stein, L. , 2001, The Distributed Annotation System *BMC Bioinformatics*, 2:7

3. Côté, R.; Jones, P.; Apweiler, R. and Hermjakob, H. , 2006, The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries, *BMC Bioinformatics*, 2006, 97

4. Ashburner, A.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M. and Sherlock, G. , 2000, Gene Ontology: tool for the unification of biology, *Nature Genetics* 25, 25 - 29

5. Shimoyama, M.; Petri, V.; Pasko, D.; Bromberg, S.; Wu, W.; Chen, J.; Nenasheva, N.; Twigger, S. and Jacob, H. , 2005, Using Multiple Ontologies to Integrate Complex Biological Data, *Bio-Ontologies SIG, ISMB 2005*

6. Plaisant, C.; Shneiderman, B.; Doan, K. and Bruns, T. , 1999, Interface and data architecture for query preview in networked information systems, *ACM Trans. Inf. Syst.*, 17 , 320-341

7. Smith, B.; Ceusters, W.; Klagges, B.; Köhler, J.; Kumar, A.; Lomax, J.; Mungall, C.; Neuhaus, F.; Rector, A. and Rosse, C. , 2005, Relations in biomedical ontologies, *Genome Biology*, 6:R46

8. Pirolli, P. and Card, S. , 1999, Information foraging, *Psychological Review*, 106 , 643-67

9. Pirolli, P.; Card, S.K. and Wege, M.M.V.D. , 2003, The effects of information scent on visual search in the hyperbolic tree browser, *ACM Trans. Comput.-Hum. Interact.*, ACM Press, 10 , 20-53

10. Bodenreider, O. and Stevens, R. , 2006, Bio-ontologies: current trends and future directions, Brief Bioinform, 7 , 256-274

11. Otgaar, D.; Dominy, J.; Maclear, A.; Gamieldien, J.; Martinez, F. and Jacobson, D. , 2006, DigraBase: A Graph-theoretic Framework for Semantic Integration of Biological Data (poster), *Bio-Ontologies SIG, ISMB 2006*

Incorporating Functional Inter-relationships into Algorithms for Protein Function Prediction

Gaurav Pandey*, Vipin Kumar
Department of Computer Science and Engineering, University of Minnesota, Twin Cities
200, Union Street SE, Minneapolis, MN 55414, USA
*To whom correspondence should be addressed: gaurav@cs.umn.edu

1. INTRODUCTION

A variety of recently available high throughput data sets, such as protein-protein interaction networks, microarray data and genome sequences, offer important insights into the mechanisms leading to the accomplishment of a protein's function. However, the complexity of analyzing these data sets manually has motivated the development of numerous computational approaches for predicting protein function (1). Several of these approaches use data mining and machine learning techniques for this task, and have produced very encouraging results. For a recent comprehensive survey on this topic, see reference (2).

Commonly used data mining techniques for the task of protein function prediction consider the functional classes to be used for annotation as independent of each other. However, it is well known that a protein may perform multiple functions, which may further have significant inter-relationships when viewed as concepts in a widely accepted hierarchical organization of functional classes such as Gene Ontology (3). Traditional techniques do not handle such inter-relationships, hence by incorporating them, the performance of protein function prediction algorithms could be improved.

In this paper, we use the similarity measure defined by Lin (4) as a measure of the similarity between two functional classes, and modify the traditional *k*-nearest neighbor classification algorithm to take this similarity into account. Evaluation of the algorithm on functional classification of gene expression data indicates that the use of inter-relationships between functional classes indeed substantially improves the accuracy of the hypotheses generated by protein function prediction algorithms.

2. PROPOSED APPROACH

The traditional k-NN classifier determines the annotations of a protein by finding all abundant functional classes in its neighborhood, which is the set of k proteins nearest to p in the data set, using the formula:

$$classes(p) = \{c \mid (\sum_{p' \in nbd(p)} sim(feature(p), feature(p')) * [c \in classes(p')]) > threshold_1\}$$

Kuramochi *et al* (5) showed that this simple algorithm performed comparably to more powerful classification algorithms such as SVMs for functional classification of gene expression data. We modified the above formula as follows, to take the similarity between functional classes into account:

$$classes(p) = \{c \mid (\sum_{p' \in nbd(p)} sim(feature(p), feature(p')) * \max_{c' \in classes(p')} \{sim(c,c')\}) > threshold_2\}$$

Thus, if a protein p is strongly expected to belong to class c, but its neighborhood does not contain enough evidence for this annotation, then the above formula enables other proteins in the neighborhood to contribute to this evidence, in proportion to the similarity of its most similar class to c. This incorporation of class similarities is expected to have the advantage of improving the predictions for proteins that do not have enough evidence for annotation by a certain class as per the original function prediction hypothesis, by enabling the transfer of annotations from close proteins annotated with similar classes. Thus, this method makes the hypothesis underlying automated function prediction more flexible by using the similarity of features of two proteins to indicate functional *similarity* (6) instead of functional *equivalence*, as assumed by most current techniques.

We use the hierarchical organization of functional classes in the Gene Ontology to model the similarity of two classes (nodes) in GO using Lin's measure (4): $sim(c_1,c_2) = \frac{2 \times [\ln p_{ms}(c_1,c_2))]}{\ln p(c_1) + \ln p(c_2)}$. Here, c_1 and c_2 are functional classes in one of the GO hierarchies, $p(c)$ is the probability of a protein being annotated with class c, and $p_{ms}(c_1,c_2) = \min_{c \in S(c_1,c_2)} \{p(c)\}$, where $S(c_1,c_2)$ is the set of common ancestors of c_1 and c_2. This

measure evaluates the similarity of two nodes in a hierarchy in terms of the population of the *least* common ancestor, and is normalized to have a value in the range of [0,1]. The use of Gene Ontology to identify inter-relationships between functional classes, and the use of the above similarity measure to quantify these inter-relationships enables us to incorporate biologically significant knowledge into our function prediction algorithm, and improve the performance of previous algorithms, as detailed in the following section.

Fig 1: Original class similarity matrix derived using Lin's measure

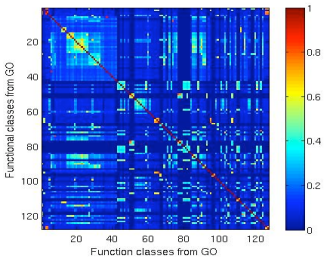


Fig 2: Optimally filtered class similarity matrix producing best AUC performance

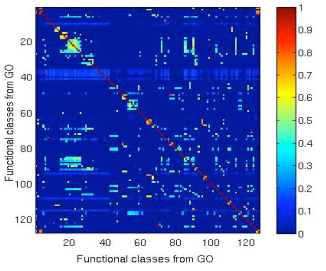
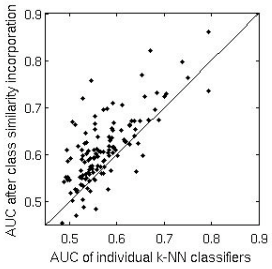


Fig 3: Improvement in AUC score of each class using class similarity



3. RESULTS AND DISCUSSION

We used Mnaimneh *et al*'s gene expression data set (7) to test the effectiveness of incorporating functional class similarities into a functional classification algorithm. This data set measures the expression of 6306 genes from *S. cerevisiae* under a set of 215 titration experiments. A set of 127 functional classes chosen were from the *biological process* ontology of GO, each having at least 10 members, and had been suggested by Myers *et al* to be testable in a wet lab (8). Figure 1 graphically shows the matrix of pair-wise similarities between these classes calculated using Lin's similarity measure. The density of this matrix suggests that it is likely to contain spurious similarities due to factors such as the abundance of classes from the *cellular process* ontology in the target set. Hence, we used an appropriate similarity filtering threshold for each class, which was done as follows. The original data set was split into two halves, and the first half was used as input to the label similarity-incorporated k-NN classifier discussed above. Through a five-fold cross validation procedure, the classification performance was evaluated for each class in terms of the AUC measure for several thresholds, and the one producing the best performance was chosen as the best filtering threshold for each class. Figure 2 shows the resultant filtered class similarity matrix.

Now, this similarity matrix is used for the predicting the functions of the genes in the other half of the original data set through a five-fold cross validation procedure, which is run multiple times in order to obtain robust estimates of the AUC scores for several classes. In addition, the basic k-NN classifier (5) is also used for predicting the functional classes of all the genes in the dataset through a five-fold cross validation procedure, and the AUC score of each class is computed. Figure 3 shows the comparison of the AUC scores of the 127 classes obtained using both the class similarity-equipped (y-axis) and the basic k-NN algorithms (x-axis). This plot shows that the performance of 102 classes is improved by considering similarities between classes, with the improvement being very significant for several classes, while the performance for the other 25 classes is only slightly deteriorated. These results show the utility of modeling similarities between functional classes as a way of incorporating the knowledge embodied in Gene Ontology, and thus producing more accurate predictions for proteins. Generalization of this concept for use with other classification methods, such as SVM, and other types of biological data, such as protein-protein interaction networks, is in progress.

4. REFERENCES

1. Marcotte, E. M., 2000, Computational genetics: finding protein function by nonhomology methods, Curr Opin Struct Biol. 10, 3, 359–365
2. Pandey, G., Kumar, V. and Steinbach, M., 2006. Computational Approaches for Protein Function Prediction: A Survey, TR 06-028, Dept of Comp Science and Engineering, University of Minnesota, Twin Cities
3. Ashburner, M. *et al*., 2000. Gene Ontology: tool for the unification of biology, *Nat. Genet.*, 25(1), 25–29.
4. Lin D., 1998, An Information-Theoretic Definition of Similarity, Proc. Intl. Conf. Machine Learning, pp 296-304
5. Kuramochi, M. *et al*, 2005. Gene Classification Using Expression Profiles: A Feasibility Study, Intl. J. Artificial Intelligence Tools, 14(4): 641-660
6. Lord PW *et al*, 2003, Semantic similarity measures as tools for exploring the gene ontology, in PSB, pp 601-612
7. Mnaimneh, S. et al, 2004, Exploration of essential gene functions via titrable promoter alleles, Cell, 118(1), 31-44
8. Myers C.L. *et al*, 2006, Finding function: evaluation methods for functional genomic data, BMC Genomics, 7:187

Integrating Sequence and Structural Biology

with the Distributed Annotation System

Andreas Prlic 1*, Thomas Down 2, Eugene Kulesha 3, Robert Finn 1, Andreas Kahari 3, Tim Hubbard 1
1 Wellcome Trust Sanger Institue, Hinxton, Cambridge, U.K.
2 Wellcome Trust/Cancer Research UK Gurdon Institute, Cambridge University, Cambridge, UK
3 European Bioinfomatics Institute (EBI) Hinxton, Cambridge, U.K
*To whom correspondence should be addressed: ap3@sanger.ac.uk

1. INTRODUCTION

New large scale techniques in biology are producing a rapidly growing amount of publicly available data. Centralized database resources are confronted with the challenges of scaling up storage facilities, managing frequent updates, and exchanging data with the community.

The Distributed Annotation System (DAS) addresses many of these issues. The DAS protocol is widely used to openly exchange biological annotations between sites distributed around the world. The problems of data distribution, handled by DAS servers, are separated from visualization and user-interface issues, which are handled by DAS clients.

DAS is a client-server system in which a client program, such as the Ensembl genome browser [1], collects and integrates information from multiple servers. It allows a single client to aggregate annotation information from multiple web servers, collate the information, and display it to the user in a single view. Little coordination is needed among the various information providers.

DAS is heavily used in the genome bioinformatics community. Recently, we have also seen growing acceptance in the protein sequence and structure communities. In particular, members of the BioSapiens project have provided a large number of DAS servers. Here we present an overview of this DAS community and present new DAS applications that we have developed recently. These include web-sites that are utilizing DAS and software applications that allow for mapping of genomic variation data onto the 3D protein structure and visualization of multiple structure alignments.

REFERENCES

[1] Ensembl 2007.

Nucleic Acids Res. Database Issue 2007
Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E.

FLORA: A novel method for predicting function from structure

Oliver Redfern*, Tim Dallman, Christine Orengo
Department of Biochemistry and Molecular Biology, University College London, London, WC1E 6BT, UK
*To whom correspondence should be addressed: ollie@biochem.ucl.ac.uk

1. INTRODUCTION

As a consequence of international genome initiatives, there are over 3 million sequences in public repositories. In addition, the Structural Genomics Initiatives (SGIs) are also increasing the rate at which new protein structures are deposited in the Protein Data Bank (PDB). Combining sequence information with three-dimensional (3D) structure data can help to identify not only the residues that are important for catalysis and molecular interactions, but also the significance of their conformation in 3D space.

Prior to the advent of the SGIs, it was uncommon for the structure of a protein of unknown function to be solved. Indeed, the motivation for its structural characterization was to shed light on its interactions and regulation. By contrast, proteins targeted by the SGIs are often selected on the basis of their lack of sequence homology to existing structures and thus could potentially also have novel functions. However, currently around 90% of the structures solved by the SGIs comprise domains that are structurally similar to those already classified in SCOP[1] or CATH[2] and therefore methods that exploit the structural data to predict function are becoming increasingly valuable.

The first step for a query structure of unknown function is frequently to search for homologues using sequence methods. By focusing on profiles of conserved residues, methods such as PSI-BLAST and Hidden Markov Models (HMMs) are able to detect distant relatives well below the twilight zone (< 35% sequence identity). However, simply because two proteins are evolutionarily related does not mean they perform similar functions.

Protein structure tends to be more conserved than sequence through evolution[3] and global structure comparison methods (DALI[4], CE[5] and SSAP[6]) can also be exploited to detect structural relatives with potentially relevant functional annotation. To aid this process, domain databases such as CATH[2] and SCOP[1] deconstruct protein chains into their constituent domains and classify these into superfamilies. However, there are numerous examples of superfamilies that exhibit a diverse range of enzymatic functions, such as the P-loop hydrolases,

To complement global homology, many groups have focused on matching structural templates containing a subset of residues that are thought to be important for function (such as the Catalytic Site Atlas, CSA[7]). However, it is often difficult to accurately score the significance of matching small structural motifs to a query structure. As a consequence, the GASP[8] method of Babbitt et al. used a genetic algorithm to design small templates based on their ability to discriminate true functional relatives from a background of SCOP domains. The results were promising, but it was only tested on 5 functionally homogenous superfamilies.

Here we introduce a novel algorithm FLORA, in addition to a benchmark data set to compare a range of current structure-based function prediction methods. A non-redundant (<35% sequence identical) set of domains from the CATH database were clustered into enzyme families based on their E.C. annotations and manual validation using the literature. 14 functionally diverse superfamilies were identified, containing more than one enzyme family (at the level of the 3rd E.C. number).

FLORA aims to use information from multiple structure alignments of each enzyme family to select residues associated with function. Patterns of sequence conservation and solvent accessibility are analyzed to locate the functional site for each family and structurally conserved residues from this local environment are selected to build a template. The predicted site was shown to coincide with known catalytic residues in the CSA in 80% of families. A graph-matching algorithm was implemented to compare enzyme templates against structures within the same CATH superfamily to identify functional relatives. Preliminary trials

have shown promising results with more than 90% of the query structures matching the correct, functionally related homologue as the top rank in the search.

The performance of FLORA has been compared to global homology methods (SSAP, CE and PSI-BLAST) as well as other published local template methods. In addition, the effect of generating global templates that were not limited to the predicted functional site was also explored.

4. REFERENCES

1. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247(4), 536-540
2. Todd, A. E., Marsden, R. L., Thornton, J. M., and Orengo, C. A. 2005 Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.* 20;348(5), 1235-1260 (2005).
3. Chothia, C. and Lesk, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5(4), 823-826
4. Holm, L. and Sander, C. 1993 Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**(1), 123-138
5. Shindyalov, I. N. and Bourne, P. E. 1998 Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11(9), 739-747
6. Taylor, W. R. and Orengo, C. A. 1989 Protein structure alignment. *J. Mol. Biol.* 208(1), 1-22
7. Orengo, C. A.*, et al.* 1997 CATH--a hierarchic classification of protein domain structures. *Structure.* **5**(8), 1093-1108
8. Porter, C. T., Bartlett, G. J., and Thornton, J. M. 2004 he Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* 32(Database issue), D129-D133
9. Polacco, B. J. and Babbitt, P. C 2006 Automated discovery of 3D motifs for protein function annotation. *Bioinformatics.* 22(6), 723-730

mGene: A Novel Discriminative *Ab-initio* Gene Finding System

G. Schweikert[+,1,2,3], G. Zeller[+,1,3], A. Zien[+,1,2], C.S. Ong[1,2], F. de Bona[1], S. Sonnenburg[4], P. Philips[1], G. Rätsch[*,1]

[1] Friedrich Miescher Laboratory of the Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany
[2] Max Planck Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany
[3] Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany
[4] Fraunhofer FIRST, Kekulèstr. 7, 12489 Berlin, Germany
[*]To whom correspondence should be addressed: Gunnar.Raetsch@tuebingen.mpg.de
[+] Contributed equally to this work

1. INTRODUCTION

As an increasingly large number of genomes are being sequenced and have to be annotated, the computational problem of gene finding has never been more important. We present a novel discriminative machine learning technique to solve the structured output learning task posed by computational gene finding. Our new gene prediction system, called *mGene* (http://www.mgene.org), provides very accurate *ab initio* gene predictions including alternative isoforms and is also able to take genomic conservation or known transcript sequences into account.

Here we build on our previous work on (a) sequence classification with Support Vector Machines (SVMs) [1,4,5] and (b) on an accurate system to predict the splice form of a gene, called *mSplicer* [3]. SVMs employ kernels for comparison of the objects to be classified. For accurate recognition of intrinsic signals such as splice sites, transcription/translation starts/stops the so-called *Weighted Degree* kernel [1,6] is particularly well suited. Using the area under the precision-recall curve as a measure of prediction performance, we obtain ~95% for *C. elegans* and ~44% for human donor splice site recognition with SVMs. This is a significant improvement over position weight matrices that achieve only ~88% and ~5%, respectively [2]. Based on these prediction methods, we demonstrated in [3] that the genome annotation of *C. elegans* can be greatly improved using novel discriminative machine learning techniques related to generalized hidden Markov models. Given the start and end of a gene's coding sequence, *mSplicer* correctly identifies all exons and introns for 95% of all genes. In *mGene* this approach was extended to a full gene finding system where the start and end of the transcript and the coding region are unknown and have to be predicted in addition to exon/intron boundaries.

2. METHODS

In a first step, we identify various signal sequences such as transcription and translation starts and stops as well as splice sites. As these signals correspond to transitions between gene segments, their proper recognition is the basis of accurate gene prediction. For each signal we independently train SVM classifiers employing a combination of string kernels individually tailored to the signal at hand [2,4,6]. We use our freely available Shogun toolbox (http://www.shogun-toolbox.org) which allows us to train on millions of examples.

In a second step, we approach the gene structure prediction problem with an extension of the label sequence learning algorithm used in *mSplicer*. As input it takes the predictions from the first step indicating possible transitions between segments. Together with additional content information (e.g. coding potential or segment lengths) signal predictions are combined in order to learn globally optimal gene structures. Using discriminative learning techniques throughout, we enforce a large margin between the score of the true gene structure and the scores of all other wrong structures (cf. Fig. 1). This approach is in contrast to many popular gene finders that fit a generative model to the data instead of learning to discriminate between correct and incorrect gene structures, which is arguably an easier learning problem. An important advantage of our method is that additional features like sequence conservation or expression measurements can be quite easily included into the method without the need of incorporating them in a probabilistic model.

3. PARTICIPATION IN THE nGASP COMPETITION

With *mGene* we participated in the nGASP genome annotation competition on nematode genomes (http://www.wormbase.org/wiki/index.php/Gene_Prediction). While the official evaluation of the predictions will be announced in the near future (expected on 15 May), we made a preliminary assessment

comparing the public genome annotation to our predictions and those of our competitors. We found considerable differences between the accuracies of the methods and a large disagreement with currently unconfirmed gene annotations. Considering experimentally confirmed genes as a ground truth, we evaluated submissions in each of three submission categories (*ab initio*; using conservation; using known sequences) on nucleotide, exon and transcript level accuracy. According to seven out of these nine criteria, *mGene* performed better than any of the 15 other methods including *Augustus* [7] and *Craig* [8].

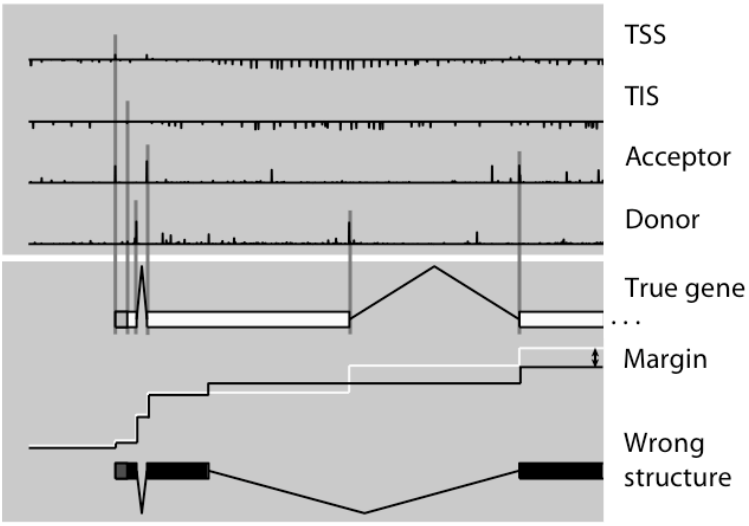


Fig. 1
Gene prediction is decoupled into two steps. In the first one (upper panel) various signal sensors are trained. Their outputs (black lines) are combined in the second step (lower panel) such that a large margin between the global score of the true gene structure (in white) and the score of any other wrong gene structure (example in black) is achieved.

4. CONCLUSION AND FUTURE WORK

The performance of our system illustrates that modern machine learning techniques hold great promise for deciphering genome sequences. So far we only demonstrated that our approach works well for *C. elegans*, but we are currently working on extending and applying the method to other genomes, including human.

In contrast to most other gene finders, *mGene* is already able to predict multiple isoforms generated by alternative splicing: Following gene finding, alternative splicing events are detected with a separate classifier [6,9] and added to the initial gene predictions (also submitted to nGASP; evaluation pending). Moreover, with our approach we can also directly include features relevant for the prediction of alternative splicing (e.g. [9,10]). By additionally extending *mGene's* structure learning algorithm (second step), we currently work towards a system that is able to directly predict splicegraphs representing multiple isoforms generated by alternative splicing.

REFERENCES

1. G. Rätsch and S. Sonnenburg. *Accurate splice site prediction for C. elegans*. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, Kernel Methods in Computational Biology. MIT Press, 2004.
2. S. Sonnenburg, P. Philips, G. Schweikert, and G. Rätsch, *Accurate Splice Site Prediction.* Submitted for publication in **BMC Bioinformatics**, April 2007.
3. G. Rätsch et al., *Improving the C. elegans genome annotation using machine learning*. **PLoS Computational Biology**, **3**(2):e20, 2007.
4. S. Sonnenburg, A. Zien, and G. Rätsch. *ARTS: Accurate Recognition of Transcription Starts in Human*. **Bioinformatics**, **22**(14):e472–e480, 2006.
5. S. Sonnenburg, G. Rätsch, and K. Rieck. *Large scale learning with string kernels*. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, Large Scale Kernel Machines. MIT Press, 2007. In press.
6. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. *Large scale multiple kernel learning*. **J. Mach. Learn. Res.**, **7**:1531–1565, 2006.
7. M. Stanke, A. Tzvetkova, B. Morgenstern, *AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome*. **BMC Genome Biology**, 7 (Suppl 1):S11. 2006.
8. A. Bernal, K. Crammer, A. Hatzigeorgiou, F. Pereira, *Global Discriminative Learning for Higher-Accuracy Computational Gene Prediction*. **PLoS Comput Biol.**, **3**(3):e54, 2007.
9. G. Rätsch, S. Sonnenburg and B. Schölkopf. *RASE: Recognition of Alternatively Spliced Exons in C. elegans*. **Bioinformatics**, **21**(Suppl.1):i369-i377, 2005.
10. R. Sorek, R. Shemesh, Y. Cohen, O. Basechess, G. Ast, R. Shamir, *A non-EST-based method for exon-skipping prediction.* **Genome Res.**, **14**(8):1617-23, 2004.

Predicting RNA-binding Proteins from their Three Dimensional Structure

Shula Shazman and Yael Mandel-Gutfreund*
Faculty of Biology, Technion- Israel Institute of Technology , Haifa 32000, Israel
*To whom correspondence should be addressed: yaelmg@tx.technion.ac.il

1. INTRODUCTION

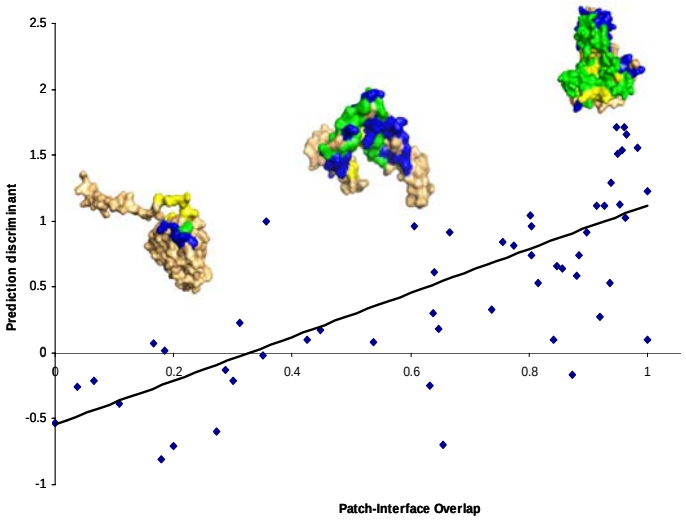
RNA has gained much interest in the recent years, mainly as a result of the large genome sequencing and the discovery of many non-coding functional RNA. Following the structure genomic initiative we expect that many proteins with potential RNA-binding function will be solved. In this study we apply a machine learning approach (Support Vector Machine) to predict RNA binding proteins from their three dimensional structure, concentrating on specific features which are characteristic of RNA binding proteins. The method is based on extracting positive electrostatic patches on protein surfaces (1, 2). Using features extracted from the positive electrostatic patches on RNA and non nucleic-acid binding proteins we trained an SVM to recognize specifically RNA-binding proteins. Applying a hold out (jackknife) approach we show that we can successfully distinguish the RNA-binding proteins from other proteins which do not bind nucleic-acids (88% accuracy). We show that our method is applicable for predicting RNA-binding proteins which are generally very diverse in terms of their structure, function and RNA recognition motifs. Moreover, as the method does not rely on sequence or structure conservation, we suggest that it could be applied to predict other RNA-binding proteins with novel folds and/or unique binding motifs. The method was further tested on an independent set of hypothetical proteins from the PDB database and successfully predicted 56 of 85 hypothetical proteins annotated by GO terms as RNA-binding proteins.

Furthermore, applying a multi class SVM we show that we can classify the proteins based on their RNA target. The multi-class SVM approach (known also as the One Versus All approach) is generally a series of binary SVM classifiers where in each classifier the members of one sub-class is separated from the rest of the data. We built three new SVM classifiers; tRNA vs All, rRNA vs All, mRNA vs All. To predict whether a given protein belongs to a specific subset we tested each of the RNA-binding proteins in our training set (46 rRNA-binding proteins, 13 tRNA binding proteins and 25 mRNA binding proteins) on the three classifiers (in each case the tested protein was held out from the training set). Finally a protein was assigned a value based on the classifier in which it achieved the highest positive discriminating value. The results of the multi-SVM test are summarized in Table 1. As shown in the Table in all three sub classes the majority of proteins were correctly assigned to their subgroup. The best results were obtained for the tRNA binding proteins where 13 of the 13 tRNA-binding proteins were clearly assigned as tRNA-binding. In the case of the rRNA binding proteins, though the majority of the proteins (74%) were given a highest score in the rRNA classifier still some protein were miss-classified mainly as mRNA binding. In the case of mRNA subclass, predictions were the weakest and only 52% of the proteins were assigned correctly. The latter results are probably due to the diversity within the mRNA subclass.

As noted, the basic assumption behind our prediction algorithm was that the electrostatic patch is reminiscent of the binding interface. Thus it is expected that the success of the method would depend on the correlation between the patch residues (identified automatically by our algorithm) and the experimentally defined RNA-binding interfaces. We tested the correlation between the patch–interface overlap and the

confidence of the RNA-binding prediction, as derived from the SVM. As illustrated in Figure 1, we found significant positive correlation (r=0.72, p<0.0001) between the percent overlap of the positive electrostatic patch and RNA-protein interface and the discriminating value obtained by the SVM. Interestingly, we found that the overlap between the positive patch and the RNA-protein interface was overall lower than for DNA-binding proteins with a larger degree of variance. This result is consistent with our previous suggestion that different than in DNA-binding proteins the largest positive patch doesn't always coincide with the RNA-protein interface. Based on these results we have added other features extracted from additional electrostatic patches on the protein surface. Nevertheless, applying a feature selection procedure we show that the properties of the largest patch still contribute mostly to the predicting power. To our knowledge our algorithm is currently the most accurate method for predicting RNA-binding function from structure.

2. FIGURES



3. TABLES

	mRNA pred	**rRNA pred**	**tRNA pred**	**Total**
tRNA	0	0	**13**	13
rRNA	8	**34**	4	46
mRNA	**13**	7	5	25

Table 2: A table summarizing the multi-SVM results for 3 subclasses of RNA-binding proteins. As shown, the majority of proteins were correctly assigned to their subgroup**.**

Figure 1 above: The correlation between the patch-interface overlap and the discriminate value obtained from the SVM classifier. As illustrated the prediction power of the algorithm depends on the success in identification of the functional interface. The correlation is exemplified on three proteins, the blue region represents the largest positive patch, yellow is the real binding interface and green denotes the overlap between the extracted patch and the real interface.
left-a false negative prediction corresponds to no overlap , middle-borderline prediction correspond to low overlap , right-strong prediction correspond

4. REFERENCES

1. Stawiski, E. W., Gregoret, L. M. & Mandel-Gutfreund, Y. 2003. Annotating nucleic acid-binding function based on protein structure. *J Mol Biol* 326, 1065-79

2. Shazman S., Celniker G., Haber O., Glaser F. and Mandel-Gutfreund Y. 2007 . Patch Finder Plus (PFplus): A web server for extracting and displaying positive electrostatic patches on protein surfaces. *Nucleic Acid Research* ,in press

3D2EC: Predicting enzyme function from structure

Marcin von Grotthuss*, Dariusz Plewczynski, and Leszek Rychlewski
BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland
*To whom correspondence should be addressed: mvg@bioinfo.pl

1. INTRODUCTION

Until not so long ago a protein structure was select to be solved only after its biochemical function was known beforehand, and whose three-dimensional structure might provide an understanding of the exact details of the mechanism underlying that function[1]. But now, when over 30 structural genomics centers have been established worldwide with the common goal of large scale, high throughput structure determination using X ray crystallography and NMR[2], the selection criteria has been changed. Currently, decisive factors mostly include attributes such as the likelihood of the protein having a new fold, or whether the protein is a representative of a large, uncharacterized protein family, or perhaps is a sole member of a novel family[3]. Function rather rarely enters the equation, and so many of the structures deposited at the PDB by the various structural genomics initiatives are of 'hypothetical proteins', i.e. proteins of unknown, or uncertain, function[4].
However, knowing the three-dimensional structure of a protein opens up the possibility of ascertaining its function from an analysis of that structure. There are a number of methods in current use for predicting function from three-dimensional structure, like Dali[5] and VAST[6] which are based on simple idea which assumes that the more similar the structure, the more similar the function is likely to be. It is well recognized that none of these structure-based methods can expect to be successful in all cases. Methods that are able to detect catalytic residues in a three-dimensional structure will give no useful information if the protein in question is not an enzyme. Also, global comparison methods can sometimes give inaccurate results, especially when a function of the query protein significantly altered during evolution while its fold remained largely unchanged (e.g. the TIM-barrel fold which supports over 60 different functions).
One way to overcome this problem is to use as many methods as possible in order to increase the chances of obtaining a helpful match. Currently, there are only two web-servers (ProKnow[7] and ProFunc[8]) available on the Internet which combine results from different methods and try to find the consensus. Both these tools almost always provide correct function of the query protein in the hit list but the consensus answer is still too often incorrect. For example, ProKnow error varies from 11% up to 60% with 93% coverage of 1507 distinct folded protein [7]. Continuing to use such error prone procedures could lead to an unmanageable propagation of errors that might at the end jeopardize progress in Biology.

2. PROTOCOL

Probably, the huge error rate is a consequence of using one averaged set of parameters, like similarity cutoff, that were optimized for the product of sensitivity and specificity. The algorithm presented here, not only combines results from 2 different methods (namely 3D-Hit and 3D-Fun) but also uses fold-specific score cutoff for both of them. The protocol was evaluated to predict the EC number of an enzyme but it can be extended to any other protein function.
The first method simply scans using the 3D-Hit program [9] a sequentially non redundant database of structures that are characterized by four cutoff values. Each value is defined by the highest, known score of structural similarity to any protein with different enzyme function at the corresponding or lower EC level. In the 3D-Hit strategy, the EC number of the protein with the strongest structural similarity is completely (or partially) assigned to the query, if the similarity score is greater than all (or any) of the cutoff values. As an example; let us consider a query protein which has the 3D-Hit score =150 to the enzyme with the EC number 1.2.3.4 and the cutoff values =100, 120, 180, 200, respectively. This structure will obtain an EC number assignment of 1.2.?.?.
All structural similarity scores are used for annotation in the 3D-Fun strategy. First, the query structure and all sequentially non redundant proteins are hierarchically clustered (grouped) by structural similarity using complete link algorithm [10], [11]. Next, the EC number is completely (or partially) assigned to each group in each clustering iteration, if all of the enzymes in the group have the same function at all (or any) of the EC levels; otherwise the EC number is assigned as unknown. As an example let us consider a cluster that contains 4 structures: the query protein and 3 enzymes with EC numbers 1.2.3.4, 1.2.3.6, and 1.2.4.1. This

cluster will obtain an EC number assignment of 1.2.?.?. For the final prediction, the enzymatic function of the smallest cluster which contains the query structure is used. In the contrary to the 3D-Hit strategy, the 3D-Fun algorithm takes into account the enzymatic function of all structures that have greater values of similarity to the query than to all other proteins of the whole set.

3. APPLICATION

We used both presented algorithms to infer the EC number for 2000 proteins from structural genomics that are currently available and have functions marked as "unknown" in the PDB file (all predictions are available at http://bioinfo.pl/PDB-UF web-page). Unfortunately (or fortunately, depending on the point of view), structural genomics initiatives tend to target structures that are less typical of the PDB as a whole and so the cutoffs derived from the whole PDB may not be entirely applicable. Therefore, we analyzed more than 100 structures with predicted EC numbers, which were recently published and functionally annotated. So far, we found only 3 incorrect assignments (that were manually corrected). However, as more structures are solved in the Protein Data Bank, the protocol will be more and more accurate and human intervention will not be required.

3. CONCLUSION

The result shows that the false positive error rate can be significantly decreased (almost to zero) by using fold-specific cutoffs. This is with agreement with two complementary rules that tell if protein function is determined by different protein folds it is not determined by fold which supports multiple functions; and if protein fold supports multiple functions it does not support function which is determined by multiple folds [12]. It means, for example, that low structural similarity (DALI Z-score=3) can be enough for correctly assigning function to some structures (like methyltransferase with a deep trefoil knot [13]) while for others, even very high structural similarity (DALI Z-score>30) could cause incorrect prediction (as in the case of the mentioned proteins with the TIM-barrel fold).

4. REFERENCES

1. Laskowski RA, Watson JD, Thornton JM: From protein structure to biochemical function? *J Struct Funct Genomics* 2003, 4(2-3):167-177.
2. Chen L, Oughtred R, Berman HM, Westbrook J: TargetDB: a target registration database for structural genomics projects. *Bioinformatics* 2004, 20(16):2860-2862.
3. Laskowski RA, Watson JD, Thornton JM: Protein function prediction using local 3D templates. *J Mol Biol* 2005, 351(3):614-626.
4. Berman HM, Westbrook JD: The impact of structural genomics on the protein data bank. *Am J Pharmacogenomics* 2004, 4(4):247-252.
5. Holm L, Sander C: Dali: a network tool for protein structure comparison. *Trends Biochem Sci* 1995, 20(11):478-480.
6. Gibrat JF, Madej T, Bryant SH: Surprising similarities in structure comparison. *Curr Opin Struct Biol* 1996, 6(3):377-385.
7. Pal D, Eisenberg D: Inference of protein function from protein structure. *Structure* 2005, 13(1):121-130.
8. Laskowski RA, Watson JD, Thornton JM: ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* 2005, 33(Web Server issue):W89-93.
9. Plewczynski D, Pas J, von Grotthuss M, Rychlewski L: 3D-Hit: fast structural comparison of proteins. *Appl Bioinformatics* 2002, 1(4):223-225.
10. Defays D: An Efficient Algorithm for a Complete Link Method. *The Computer Journal* 1977, 20:364-366.
11. Murtagh F: A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal* 1983, 26:354-359.
12. George RA, Spriggs RV, Thornton JM, Al-Lazikani B, Swindells MB: SCOPEC: a database of protein catalytic domains. *Bioinformatics* 2004, 20 Suppl 1:I130-I136.
13. von Grotthuss M, Plewczynski D, Ginalski K, Rychlewski L, Shakhnovich EI: PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics. *BMC Bioinformatics* 2006, 7(1):53.

ConFunc: Feature derived Profiles for Functional Annotation

Mark N Wass & Michael J E Sternberg*
Structural bioinformatics Group, Division of Molecular Biosciences, Imperial College London,
South Kensington, London, SW7 2AZ, UK
*To whom correspondence should be addressed: m.sternberg@imperial.ac.uk

1. INTRODUCTION

Protein functional annotation is an important task of the genomics era. The ability to obtain rapidly protein sequences and sequence genomes has resulted in many proteins whose function has not been experimentally characterised. Further this characterisation process is slow compared to sequencing itself resulting in a need for approaches to predict protein function in order to obtain accurate annotations. The number of sequences requiring annotation makes it important that such methods are automated enabling them to annotate whole genomes without human intervention.

We have previously demonstrated the ability to use Position Specific Scoring Matrices (PSSMs) to predict protein molecular function using Gene Ontology (GO). Phunctioner (1) uses protein structural alignments for which PSSMs are generated for each potential GO term present among the initial protein structures used. A query protein is then scored against each PSSM to predict its function. As Phunctioner relies upon structural alignments it is limited to structural space. Here we demonstrate a general approach, ConFunc, similar to Phunctioner that is applicable to the much more extensive sequence space and could prove an effective tool for genome annotation. ConFunc is available for use as a web server at http://www.sbg.bio.ic.ac.uk/confunc.

2. METHOD

ConFunc uses GO to direct the function prediction process, by splitting sets of sequences identified by PSI-BLAST (2) into sub-alignments according to their GO annotations. Each GO term sub-alignment is then used to identify conserved residues within that group, for which a PSSM profile is generated. This combination of steps produces a set of feature (i.e. GO annotation) derived profiles from which protein function is predicted. Many methods that predict functional residues use phylogenetics approaches (3-5) to group homologous sequences. The power of ConFunc is that the grouping of sequences by GO annotation not only enables the identification of critical residues associated with a particular function but further enables them to then predict protein function. The use of GO makes it possible to predict a full range of protein functions and is not limited to enzyme function.

3. RESULTS

The direct transfer of protein function from one homologue to another is ineffective at low levels of sequence identity (6-9) making it important that alternative methods can perform well in such cases. The performance of ConFunc has therefore been benchmarked for a non redundant set of GO annotated protein sequences from Swiss-Prot, where homologues above 30% sequence identity have been removed to simulate this scenario. ConFunc performance is compared with the predictions of annotation transfer from the top BLAST and PSI-BLAST hits.

Performance of each of the methods in the benchmark is assessed over all levels of Gene Ontology using the measures of recall and precision as defined below

$$\mathrm{Re}\,call = \frac{TP}{N_A} \qquad \mathrm{Pr}\,ecision = TP \Big/ (TP + FP)$$

where *TP* and *FP* are the total number of true positive and false positive predictions respectively and N_A is the total number of annotations in the test set. The results are displayed as Precision-Recall graphs (Figure 1) which demonstrate that ConFunc is able to obtain greater recall and precision than both BLAST and PSI-BLAST. ConFunc obtains a coverage of 67% compared to a maximum coverage of 86% for BLAST and 100% for PSI-BLAST. This occurs because ConFunc requires multiple

annotated homologues to be able to generate GO term PSSMs to infer function whereas annotation transfer by BLAST and PSI-BLAST only require a single annotated sequence. We have therefore considered the performance of the methods only for sequences where they all make predictions (Figure 1b), which show that where ConFunc makes predictions, it can obtain good levels of recall with high precision, outperforming both BLAST and PSI-BLAST. The results demonstrate the effective use of protein sequence to identify conserved functional residues and in turn use them to predict protein function.

4. FIGURES

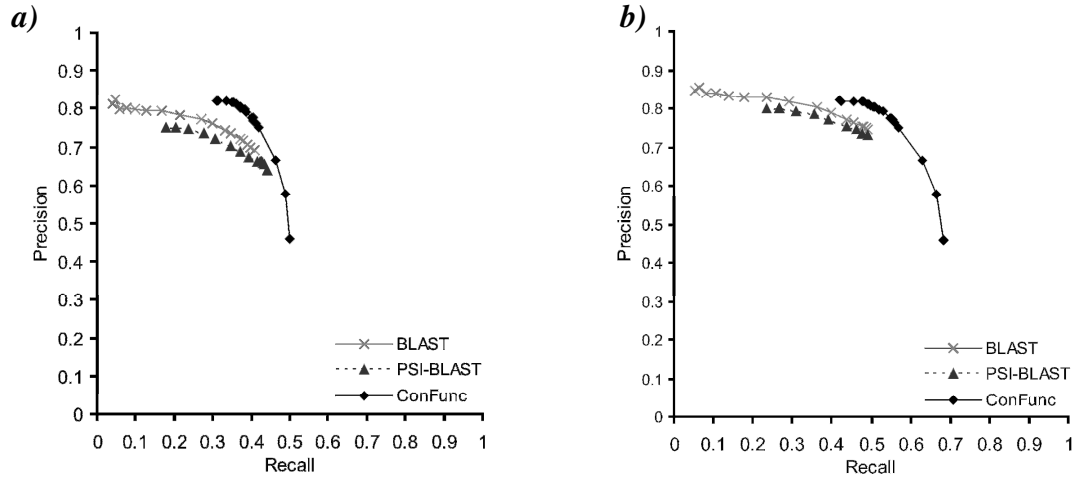


Figure 1. Assessing ConFunc Function Prediction. Precision-Recall graphs are shown for ConFunc, BLAST and PSI-BLAST function prediction. a) Results for complete protein sequence test set. b) Results for sequences in the test set where all three methods make a prediction.

5. REFERENCES

1. Pazos, F. and Sternberg, M.J. 2004. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A*, 101:14754-14759.

2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389-3402.

3. Lichtarge, O., Bourne, H.R. and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257:342-358.

4. Berezin, C., Glaser, F., Rosenberg, J., Paz, I., Pupko, T., Fariselli, P., Casadio, R. and Ben-Tal, N. 2004. ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics*, 20:322-1324.

5. Aloy, P., Querol, E., Aviles, F.X. and Sternberg, M.J. 2001. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol*, 311:395-408.

6. Hegyi, H. and Gerstein, M. 2001. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res*, 11:632-1640.

7. Devos, D. and Valencia, A. 2000. Practical limits of function prediction. *Proteins*, 41:98-107.

8. Todd, A.E., Orengo, C.A. and Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*, 307:1113-1143.

9. Rost, B. 2002. Enzyme function less conserved than anticipated. *J Mol Biol*, 318:595-608.

Molecular Function Prediction based on Local Function Conservation in Sequence and Structure Space

Nils Weinhold, Oliver Sander, Francisco S. Domingues, Thomas Lengauer, Ingolf Sommer*
Max Planck Institute for Informatics, Stuhlsatzenhausweg 85, 66123 Saarbruecken, Germany
*To whom correspondence should be addressed: sommer@mpi-sb.mpg.de

1. INTRODUCTION

While protein sequence and structure databases grow at a rapid rate, a huge amount of proteins remains without functional annotations, motivating the development of accurate methods for automatic function prediction. Here we propose the Godot method for predicting molecular function given sequence and structure of a protein. The method is based on a new concept called functionally conserved regions. A region is defined as the 200 nearest neighbors of a protein with respect to sequence and structure similarity. A functionally conserved region is characterized by very similar proteins exhibiting identical functions. When predicting functions for an uncharacterized query protein, the method locates the nearest neighbors of the query and produces a ranking of GO terms based on the distances to the nearest neighbors and their respective functionally conserved regions.

2. METHOD

The Godot method, as outlined in Figure 1, can be roughly split into a training phase performed only once (A) and a prediction phase performed for each query structure (B). During training, 3499 protein domains from a representative set provided by the ASTRAL Compendium (1) are compared against each other using different similarity measures (A1). Currently, we use four similarity measures - two sequence based and two structure based, namely local profile alignment, global profile alignment, CE (2) and TM-Align (3). The result of each comparison is a similarity matrix (A2), which can be regarded as a distance measure in protein space. For every protein domain in the dataset (or point in space) we determine the local conservation of its annotated molecular functions. Based on the presence and absence of these functions in the local environment of the protein space, we use logistic regression (A3) to represent the degree of conservation. As a result we obtain one logistic curve for each pair of protein domain and function. In the prediction phase, an uncharacterized query protein is compared to all protein domains in the dataset (B1). Based on the nearest neighbors and their similarity to the query, we estimate raw function conservation scores for all GO terms annotated to the neighbors (B2). The raw function conservation scores are then weighted along the GO graph structure, yielding a cumulated function conservation score for each GO term (B3). Thus, for a query we obtain a list of GO terms ranked by cumulated function conservation scores. The representative set of protein domains is annotated with 1806 GO terms, which is a coverage of 23.8% of the currently 7581 molecular function terms in the GO.

3. RESULTS

We have assessed the method's performance by nested cross-validation, visualizing the outcome using precision-recall graphs. In Figure 2, the cross-validated performance of the Godot method is shown, compared to a baseline predictor. The baseline background-frequency predictor outputs a list of GO terms sorted according to their frequencies observed in the dataset. For both prediction methods the ranked list of GO terms is used to produce a precision-recall curve. Under cross-validation settings we obtain one such precision-recall curve for each of the 3499 protein domains. Each curve in Figure 2 displays the median precision of 3499 curves at 100 equidistant sampling points along the recall axis. An optimal predictor's curve would pass through (1,1). We observe that the background-frequency predictor is far away from this point while the Godot method comes close. In addition to the precision-recall evaluation, we benchmark Godot according to the published PHUNCTIONER protocol (4): at a specificity of 90%, Godot achieves 55% sensitivity while PHUNCTIONER achieves approximately 36% (cf. Fig2b of (4); evaluation on GO level 3).

4. DISCUSSION & CONCLUSIONS

Recent approaches to function prediction combine several protein features into so-called hybrid methods (5,6), or combine individual function predictions as meta servers (7). Using raw sequence and structural similarity our method is a hybrid. We use similarity merely as a basis for modeling the functional space by identifying functionally conserved regions. Employing function conservation scores improves automatic function prediction and yields reliability estimates for the predictions. The Godot method provides a modular framework, allowing facile extension through inclusion of other similarity measures, such as motifs or local shape properties. In addition to the trade-off between coverage and accuracy, performance significantly depends on the procedure for selecting proteins for the test set. Tested on a comprehensive set of proteins, Godot achieves considerable coverage at high accuracy.

5. OUTLOOK

The Godot method for predicting molecular function will be made publicly available as a web-server, which we will present on the conference.

6. REFERENCES

1. Chandonia, J.M., Hon G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S.E. 2004. The ASTRAL compendium in 2004. *Nucleic Acids Research 32:*D189-D192.
2. Shindyalov, I. N., Bourne, P. E. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11: 739–747.
3. Zhang Y., Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–710.
4. Pazos, F., Sternberg, M.J.E. 2004. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A*, 101(41):14754–14759.
5. Pal, D., Eisenberg, D. 2005. Inference of protein function from protein structure. *Structure*, 13:121–130.
6. Laskowski, R. A., Watson, J. D., and Thornton, J. M. 2005. Pro-Func: a server for predicting protein function from 3D structure. *Nucleic Acids Res*, 33(Web Server issue):W89–W93.
7. Friedberg, I., Harder, T., Godzik, A. 2006. JAFA: a protein function annotation meta-server. *Nucleic Acids Res*, 34(Web Server issue):W379–W381.

7. FIGURES

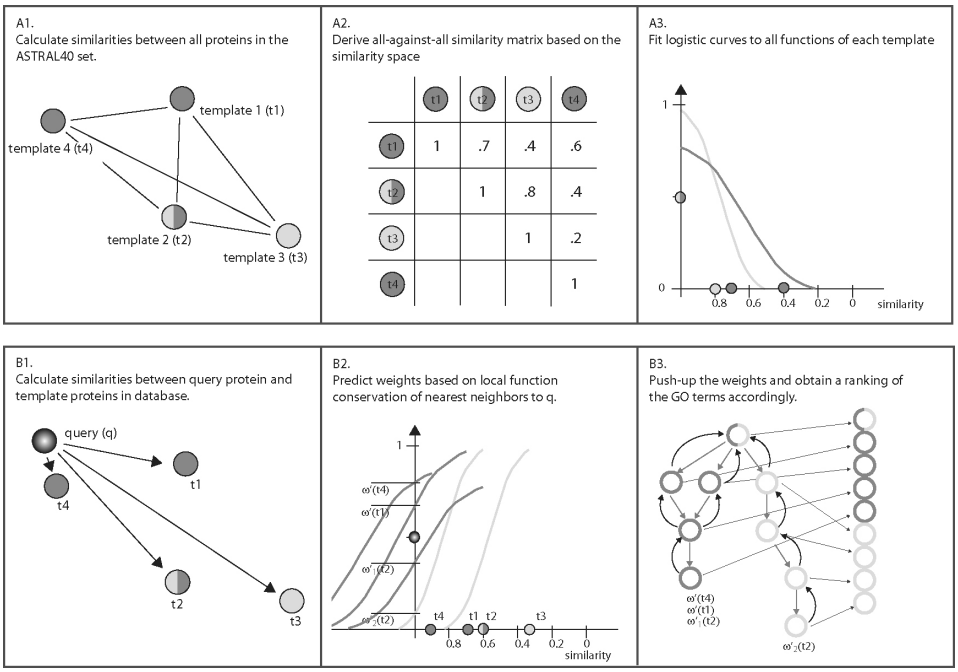


Figure 1: Overview of the Method. We exemplify the Godot method on a set of four template proteins (t1 - t4) having two different functions (drawn in light and dark grey, respectively). The training procedure (upper row) consists of similarity calculations (A1), yielding four different similarity matrices one of which is shown (A2). Based on these similarities, logistic curves are fitted for each molecular function in the dataset (A3). The prediction (lower row) constitutes similarity computations between the query protein and the proteins in our dataset (B1), which are then used to predict the conservation of molecular functions in the queries proximity (B2). The final ranking of GO terms is obtained using weighting schemes along the GO graph structure (B3). See Method section for details.

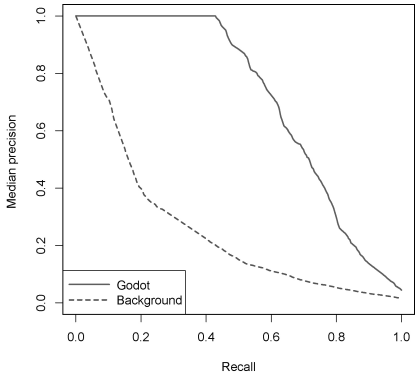


Figure 2: Test Performance of Cross-validation. The plot shows the performance of the Godot method compared to a background-frequency predictor, which predicts GO terms according to their abundance in the dataset - independently of the query structure. Each curve represents the median performance of 3499 individual predictions. The optimal performance point is in the upper right corner.

LIGHTNING TALKS AND POSTERS

GO classification of *Medicago truncatula* proteins: Using and improving SIFTER in an integrated pipeline

Anika Joecker[1]*, Barbara Engelhardt[2], Heiko Schoof[1]
1 Plant Computational Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany
2 Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California, United States of America

*To whom correspondence should be addressed: joecker@mpiz-koeln.mpg.de

1. INTRODUCTION

Annotation transfer, which confers functional annotations to a query sequence from a putative homolog, is the most common method for gene function prediction. However, this general method has many drawbacks, and makes systemic errors by not accounting for duplication events, evolutionary rate variation, and incorrect annotations. Phylogenomic analysis has been proposed to address these problems by explicitly using the evolutionary history of a gene to transfer functional annotations between homologous genes. One of the most promising tools in this area is SIFTER [1], which uses a statistical inference algorithm to propagate molecular function terms within a phylogenetic tree.

We used SIFTER to assign molecular function Gene Ontology terms to protein coding genes in the *Medicago truncatula* (Barrel Medic) genome annotation project (IMGAG). Our pipeline for this process takes the amino acid sequence of a query protein as input. We first run BLAST against proteins with an experimentally verified GO term assignment, and we extract all of the homologous genes that overlap with the query sequence by at least 70% and have an E-value less than 1. We align these proteins with MUSCLE [2] and build a phylogenetic tree using QUICKTREE [3]. We reconcile the gene trees and species trees using FORESTER [4]. The input to SIFTER is this reconciled phylogeny and all homologous genes with experimentally verified GO annotations. We assign the GO term with the highest probability as computed by SIFTER to the query protein.
Using SIFTER, we were able to predict functional annotations for 9094 proteins out of the 43616 predicted proteins in the *Medicago* genome.
We manually annotated 100 Medicago proteins, and found that, on this set of proteins, SIFTER's molecular function predictions achieved 96% accuracy versus 94% accuracy for annotation transfer using BLAST and InterProScan. In 25 cases SIFTER's prediction was more specific than the text description assigned by the most significant annotated BLAST hit against the TIGR database and the most significant hit against the InterPro database.

However, this analysis revealed some weaknesses of SIFTER. If the sequence alignment is poor and the resulting phylogenetic tree is inaccurate, or if the available GO annotations for the homologous sequences are too sparse, SIFTER can make incorrect predictions. To address these problems and to automate the functional predictions for additional genes, we modified the SIFTER pipeline in three ways.

To create a reliable alignment and reconstruct a high-quality phylogenetic tree, we implemented the following pipeline. We use iterated blast searches to identify candidate orthologs, in- and outparalogs and to compile a comprehensive phylogenetic neighborhood for the query gene.
We will implement a version of the pipeline that removes sequences from the set of homologs, that do not share domain architecture with the query protein, and compare SIFTER results between the two versions.
We align the homologous proteins with MUSCLE, and mask columns in the resulting alignment with >60% gaps. We build a phylogenetic tree using the highest quality method available given the size of the tree (i.e., neighbour joining for large trees, maximum parsimony for smaller trees, and maximum likelihood for trees with fewer than 20 proteins). As before, we reconcile the trees using FORESTER.

The version of SIFTER that we applied to this data incorporates interaction data, expression data, structure information and different ontologies, including EC numbers, MapMan bins [5], CDD and InterPro domains to compute the probabilities that each protein has a particular function. Each of these additional sources is

used to either slow down mutation (e.g. in case of similar expression profiles) or speed up mutation (e.g. in case of different interaction partners) within the SIFTER framework.
An example of this is shown in Figure 1.

We will rerun the new pipeline on all Medicago genes again and compare the results with the previous results from SIFTER and from BLAST results. Additionally we will make this method of function annotation available to the public by supporting this pipeline and the SIFTER program as BioMOBY [6] web services and as a Taverna [7] workflow.
This SIFTER pipeline will be used for GO term prediction in the tomato annotation project (ITAG).

2. FIGURES

Figure 1: Transferring of function within the phylogentic tree. If homologous proteins share the same expression profiles or have the same interaction partners, we reduce the rate of mutation within their subtree.

3. REFERENCES

1. B. E. Engelhardt et al. 2005.Protein molecular function prediction by bayesian phylogenomics. *PLoS Computational Biology* 1(5):e45

2. R. C. Edgar 2004. MUSCLE: multiple sequence alignment with high accuracy and high thoughput. *Nucleic Acid Research* 32(5):1792-1797

3. K. Howe et al. 2002. QuickTree: building huge neighbour –joining trees of protein sequences. *Bioinformatics* 18(11):1546-1547

4. C. M. Zmasek and S. R. Eddy 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17(9):821-828

5. O. Thimm et al. 2004. Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* 37(6):914-939

6. M. Wilkinson and M. Links 2002. BioMOBY: an open source biological web services proposal. *Briefings in Bioinformatics* 3(4):331-341

7. D. Hull et al. 2006. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research* 34(Web Server issue):W729-W732

Protein Function Annotation by Subsequence based Feature Map

Ö.Sinan Sarac[1], Özge Gürsoy-Yüzügüllü[2], Rengül Çetin-Atalay[2], Volkan Atalay[1]*
Dept. of Computer Engineering, Middle East Technical University, 06531 Ankara, Turkey
Dept. of Molecular Biology and Genetics, Faculty of Science, Bilkent University, 06533 Ankara, Turkey
*To whom correspondence should be addressed: volkan@ceng.metu.edu.tr

1. INTRODUCTION

Subsequence based methods provide efficient function annotations for proteins by using conserved subsequences among a class of proteins. The main idea is that, since functionally important regions (catalytic sites, binding sites, and structural motifs) are conserved over much wider taxonomic distances than the sequences themselves, conserved subsequences among different proteins are strong indicators of functional or structural similarity. Thus, in subsequence-based approach, feature vectors are constructed according to the existence of specific motifs or domains in the protein sequences. Although motifs are powerful discriminators even in low similarity (remote homology) situations, motif finding is a difficult task, especially for protein sequences since there are 20 different amino acids and many plausible mutations. One other issue is that, depending on the classification task, proteins to be classified might not have a common motif at all. As an example, in the problem of subcellular localization, when discriminating between nuclear and cytosolic proteins, it is not possible to find motifs since the proteins in the same class can have very divergent function, structure and thus, sequence.

In this study, we describe a feature space mapping, called subsequence profile map (SPMap), that takes into account the information coming from the subsequences of a protein. We decompose a given family of protein sequences into fixed-length subsequences and cluster similar subsequences. A mapping can then be defined as the distribution of the subsequences of a new protein sequence over these clusters. This approach incorporates information both coming from important subregions that are conserved over a family of proteins as well as the overall sequence similarity. Instead of focusing on the motifs, SPMap considers all of the subsequences as a distribution over a quantized space by discretizing and reducing the dimension of an otherwise huge space of all possible subsequences. SPMap improves the idea used in P2SL tool which exhibit comparable or better performance with respect to available methods for subcellular localization (1).

SPMap is a discriminative method which requires positive and negative examples to classify and annotate proteins whose functions are not known. Feature space representation of a protein sequence is the distribution of its subsequences over a map of generative models. The important step here is the clustering of subsequences. Note that the space of all possible subsequences of length l is of size 20^l, since there are 20 possible amino acids. Instead of working in this very high dimensional space, we quantized this space using the clusters of subsequences that are actually existing in the positive training examples. One should note that, as we clustered the subsequences, we were not actually looking for underlying groupings. The aim here was to generate a meaningful quantization of the subsequence space that especially represents groups of frequent and similar subsequences in the positive training data. These subsequences might have been conserved because of their importance for the function of that class of proteins and we wanted our feature space to take them into account. Clustering algorithm is similar to the average link hierarchical clustering; however it can be implemented very efficiently without calculating all the pairwise distances. Distance between a pair of subsequences is calculated by the replacement cost of individual amino acids and replacement cost was based on an amino acid similarity matrix, since it allows us to incorporate evolutionary information in finding and representing important conserved regions of a family of proteins. After the clustering step, we generated a probabilistic profile for each cluster. A probabilistic profile for a cluster is an $l*20$ matrix, where l is the length of a subsequence. An entry of this matrix represents the probability of amino acid to occur at that specific position of the subsequence. Protein sequences were represented in this feature space as the distribution of their subsequences over the generated subsequence profile map. All the subsequences of a protein were extracted to construct a feature vector. Each subsequence was then compared with each probabilistic profile and a probability was calculated. The value for a dimension of the feature vector is set to the probability of highest scoring subsequence of protein on probabilistic profile.

2. RESULTS

In all of the experiments, Blosum62 matrix was employed as the similarity matrix although it is possible to

use different similarity matrices depending on the sequence divergence or the taxonomic distance between the proteins to be classified. Length of the subsequences was set to 5. Setting the subsequence length to 5 did not mean that we sought for motifs of 5 amino acid length. In SPMap, motifs were the overall distribution of the subsequences over the profiles constructed from resulting 5 length subsequence clusters. Hence subsequence length 5 allowed us to capture longer motifs as a distribution over more than one profile. Threshold similarity value in the clustering algorithm was fixed to 8 which may allow upto approximately to 3 mutations in a 5 amino acid length subsequences using Blosum62. One of the advantages of SPMap is that it works well on wide range of different classification tasks with these default parameter values. We first performed tests on the subcellular localization dataset on which P2SL was trained and tested. In order to assess and compare the capabilities of SPMap, we then did tests on G-protein coupled receptor (GPCR) subfamily level classification that are known to be hard to classify. In addition, GPCR dataset is well studied in the literature and this allows us to compare SPMap with other available methods. On the other hand, since protein function is a vague term with many levels and aspects, and to show the capability of SPMap for functional classification problems, we performed tests on different datasets of functional aspects of proteins represented in the Gene Ontology (GO) framework.

Dataset for subcellular localization tests that we used were composed of 4 different classes: ER targeted (ER), cytoplasmic (C), mitochondrial (M) and nuclear (N) (1). In a one-versus-all setting, ROC scores were above 0.95 with a standard deviation of 0.005 for all localizations. In a second setting, classifiers for each localization were combined using the winner-take-all principle. Each test sample was assigned to the location whose classifier produced the highest SVM score. Accuracies by averaging 4-fold cross-validation tests were 88.54%, 86.96%, 83.97% and 93.48% for nuclear, cytoplasmic, mitochondrial and ER targeted proteins and these accuracies were all above those obtained by P2SL. For GPCR subfamily classification, we used the dataset presented in (2) to compare with the results of various classifiers presented in (2) and (3). Same train and test splits were used for 2-fold cross validation for fairness of comparison. Results are given in Table 1.

Table 1. Comparison of accuracy of various classifiers at GPCR level I and II subfamily classification.

Classifier	Level I accuracy (%)	Level II accuracy (%)
Fisher SVM	88.4	86.3
Blast	83.3	74.5
SAM-T2K HMM	69.9	70.0
kernNN	64.0	51.0
Decision tree	77.3	70.8
Naïve Bayes	93.0	92.4
SPMap	95.4	93.8

Furthermore, we extracted 6 families from GO database at molecular function level under metal ion binding (GO:0046872) and 10 families at the biological process level under Cellular Protein Metabolic Process (GO:0044267) level. In these experiments, we applied a filter that only allowed protein sequences annotated by one of the three evidence code, namely IC, IDA, and TAS. In addition, we omitted families that have less than 10 proteins after the filter. 4-fold cross validation was performed and the average ROC scores were between 0.51 and 0.93.

3. ACKNOWLEDGEMENT

This study is partially supported by TUBITAK under EEEAG-105E035.

4. REFERENCES

1. Atalay V. and Cetin-Atalay,R. 2005. Implicit motif distribution based hybrid computational kernel for sequence classification. *Bioinformatics* 21:1429-1436.

2. Karchin, R., Karplus, K., and Haussler, D. 2002. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 18:147–159.

3. Cheng, B.Y.M., Carbonell, J.G., and Klein-Seetharaman, J. 2005. Protein classification based on text document classification techniques. *Proteins* 58:955–970.

Heuristic rule based protein function prediction

Richard Albang[1], Eduardo Torres-Schumann[2], Dieter Voges[1], Karsten Wenger[1], Dieter Maier[1*]
1 Biomax Informatics AG, Lochhamer Str. 9, 82152 Martinsried, Germany
2 now at CIS, University of Munich, Oettingenstr. 67 80538 Munich, Germany
*To whom correspondence should be addressed: dieter.maier@biomax.com

1. INTRODUCTION

Although manual annotation covers less than 10% of the known protein sequences, it provides the gold standard for functional annotation. We have developed a heuristic rule based method for automatic function prediction that aims to capture the human decision process during manual annotation. We compare this system with two machine learning approaches using the manually annotated Saccharomyces cerevisiae genome as benchmark. Comparing predictions from decision tree, Bayesian network and heuristic rules we find that the heuristic rules provide the highest precision (86,4%) and recall (74,1%) with the given setup.

A) Heuristic rule generation: Heuristic rules were generated from self-descriptions and interviews of expert annotators at Biomax. Descriptions about how the experts evaluate sequence analysis results and create functional predictions where compared with observations of the actual annotation process. From the resulting descriptions, a set of rules was formalized in IF ... THEN form and integrated into Prolog code. We find that human annotators create annotation mostly by judging the feasibility of information transfer from sequence similarity based methods. We detected several strategies the experts use to avoid the caveats of this form of functional annotation. They minimise false predictions from incorrect homology association by integrating different parameters of the sequence similarity search, such as score or percent identity, with their knowledge about the type of protein sequence such as highly conserved ribosomal proteins or less conserved transcription factors. The inaccurate transfer of annotation from multi-domain proteins or different sub-family proteins is avoided by analysing correlation between overall sequence length and sequence similarity, consideration of domain coordinates and knowledge about protein families. Finally, transfer of inaccurate information from errors in available annotation is minimised by sampling some of the original literature and by knowledge about annotation quality within different databases. The heuristic rules therefore are generally concerned with the evaluation and integration of sequence similarity, protein family assignment and the quality of databases and database annotation. We have not attempted to include a text mining based literature sampling into this first approach of heuristic rule generation as this would have been beyond the scope of the current analysis. Expert knowledge about method parameters and idiosyncrasies as well as experience with the quality of different underlying data sources (e.g., SwissProt versus TrEMBL) and specific keywords appearing in the result (e.g., “probable”) was used to create a confidence value (c-value) for each predicted function (Figure 1).

Figure 1. Example heuristic rule.

IF Organism is “eukaryote” AND signal peptide length is below 15 **THEN** confidence is c-value = signal peptide length * 0.5
IF Organism is “eukaryote” AND signal peptide length is between 15 to 17 **THEN** confidence is c-value = 50
IF Organism is “eukaryote” AND signal peptide length is between 17 to 25 **THEN**confidence is c-value = 100

```
signalp_Setc-val00(Signal_peptide_length,SelectedOrganism,c-val) :-
SelectedOrganism == '[organism_eukaryote]', Signal_peptide_length < 15,
c-val is Signal_peptide_length * 0.5.
signalp_Setc-val00(Signal_peptide_length,SelectedOrganism,c-val) :-
SelectedOrganism == '[organism_eukaryote]', Signal_peptide_length > 14, Signal_peptide_length < 18,
c-val = 50.
signalp_Setc-val00(Signal_peptide_length,SelectedOrganism,c-val) :-
SelectedOrganism == '[organism_eukaryote]', Signal_peptide_length > 17, Signal_peptide_length < 25,
c-val = 100.
```

Rule to integrate the knowledge about SignalP prediction reliability, in IF ... THEN form and subsequent transformation into Prolog code.

All predicted functions are mapped to GO terms and the c-values for each term are normalised to a scale of 0-100. GO term c-values are integrated first within an individual method (e.g., for all *blastp* hits) and then between methods. Using the SGD Saccharomyces cerevisiae annotation(1) as Benchmark and a c-value cutoff of 50 to assign functional predictions, we reached a precision of 86,4% at a recall of 74,1%. Size and

complexity of the current rule base is quite manageable with about 200 rules forming independent rule groups around specific methods.

B) Evaluation of prediction quality: In the literature values of 60-80% recall and 70-95% precision are reported for machine learning based large scale automatic function prediction (2,3,4). However, due to the lack of common benchmark data and different evaluation metrics, it is difficult to compare these results with each other or with our results. We have therefore set up two machine learning based function prediction experiments, a decision tree and a Bayes net, using the SGD Saccharomyces cerevisiae dataset. As metric we used a very stringet measure of recall, namely the *number of correctly predicted GO terms / number of manually annotated GO terms*. According to this definition only detailed predictions equivalent to the manually annotated function will add to the recall. For decision tree learning we used the C4.5 algorithm. Following a workflow proposed earlier (5) 60% of all ORFs where used to generate the decision tree. The resulting rules were validated against another 20% of ORFs. Valid rules were then tested on the remaining 20% of ORFs. For Bayes net learning (using the R implementation), we predefined an expert derived topology and used 67% of all ORFs to learn the connecting probabilities. The conditional probability distributions were approximated by Gaussian functions, in the case of continuous variables parameterised by the mean and the standard deviation. The resulting net was tested against the remaining 33% of ORFs. To calculate a baseline prediction quality we transferred the description from either the best blastp hit (blastp best hit) or all blastp hits above an e-value cutoff of e-10 (blastp e-10) and calculated the resulting recall and precision. For the blastp results, the trade off between recall and precision becomes very obvious, varying between 30-80% recall and 60-16% precision. The decision tree method in our hand provided very specific predictions and, at 90%, the highest precision of all approaches. However due to our stringent definition, it reached only 0,6% recall, allthough a prediction was available for 22% of all ORFs. Bayes net and heuristic rules reached similar recall at 72% and 74% respectively. In part this is explained by the Bayes net topology, which has been based on heuristic knowledge as well. The precision of the heuristic rules at 86% clearly outperforms the Bayes net at 48% and is second only to the decision tree.

As detailed above, the heuristic rule based approach provides high quality prediction with good coverage. Even keeping in mind that our focus was on the heuristic rule based system, and therefore improvements to the setup of our machine learning approaches are clearly possible, the integration of detailed heuristic rules should in our oppinion, be strongly considered for automatic function prediction systems.

Table 1 Comparative prediction quality

	Heuristic rules	Decision tree	Bayes net	Blastp best hit	Blastp e-10
Recall (%)	74,1	0,6	72,4	29,7	77,4
Precision (%)	86,4	90,2	48,1	58,2	16,1

2. REFERENCES

1. Dwight S., Harris M., Dolinski K., Ball C., Binkley G., Christie K., Fisk D., Issel-Tarver L., Schroeder M., Sherlock G. et al (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Research* 30:69-72
2. Kretschmann, E., Fleischmann W., Apweiler R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* 17:920-926
3. Green M., Karp P. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5:76-92
4. Vinayagam A., Konig R., Moormann J., Schubert F., Eils R., Glatting K. and Suhai S. (2004) Applying Support Vector Machines for Gene Ontology based gene function prediction. *BMC Bioinformatics* 5:116-130
5. Clare A. and King R.(2002) Machine learning of functional class from phenotype data. *Bioinformatics* 18:160-166.

Identification and characterization of taxonomic variations of enzymes using Chisel computational environment.

A. Rodriguez, M. Syed, and **N. Maltsev**
Argonne National Laboratory, Computation Institute, the University of Chicago

Availability of large volumes of sequence and enzymatic data for taxonomically and phenotypically diverse organisms now allows for systematic exploration of the adaptive mechanisms that led to the diversification of enzymes, in terms of their kinetic and enzymatic properties, subunit composition, cofactor preferences, and other properties.

However, the characterization of the molecular variations of enzymatic functions specific to taxonomic groups and phenotypes requires a new generation of tools for high-resolution comparative and evolutionary analysis. We present our strategy for identification, characterization and evolutionary analysis of taxonomic and phenotypic variations of enzymes implemented in the Chisel system (http://compbio.mcs.anl.gov/CHISEL).

Chisel is an algorithmic approach and a computational framework, for automated and interactive identification, comparative analysis, and characterization of evolutionary variations of enzymes. Chisel contains 8,575 clusters and subsequent computational models specific for 939 distinct enzymatic functions and, when data is sufficient, their taxonomic variations.
The system includes the following components:

a. an Enzymatic Knowledge Base (EKB) – providing a basis for the hierarchical clustering of the enzymatic sequences and their annotation.
b. a hierarchical rules-based clustering algorithm for identification of enzymatic functions and their taxonomic and phenotypic variations;
c. a library of computational models for classification of un-annotated sequences;
d. a Web-based user interface with a suite of tools for interactive identification, comparative, evolutionary analysis, and annotation of the enzymatic sequences by expert users.

Chisel Algorithm. The Chisel algorithm performs rules-based clustering of initial seed sets of homologous sequences (e.g. homologous sequences provided by users, PIR superfamilies) into clusters based on similarity and available sequence features (e.g. domain composition, existence of transmembrane helicies) and others. The resulting clusters are function-specific and, in some cases, function and taxonomy specific (*i.e.*, contain sequences performing the same enzymatic function and originating from the same taxonomic group of organisms). The automatically generated models corresponding to these clusters (HMM and PSSM profiles, multiple sequence alignments, consensus sequences) can be used directly for classification of un-annotated sequences by using the Chisel classification tools. The automatically generated Chisel clusters and their subsequent computational models may be interactively refined by an expert using the additional Chisel tools.
Validation of the Chisel clusters was done using the jackknife approach (Zhang and Chou, 1995). Testing was performed with a total of 19,905 experimentally verified protein sequences (annotated with experimental GO evidence codes and extracted from references in the BRENDA database).

Fig. 1 presents an example of the analysis of the Zn-containing alcohol dehydrogenases using Chisel.

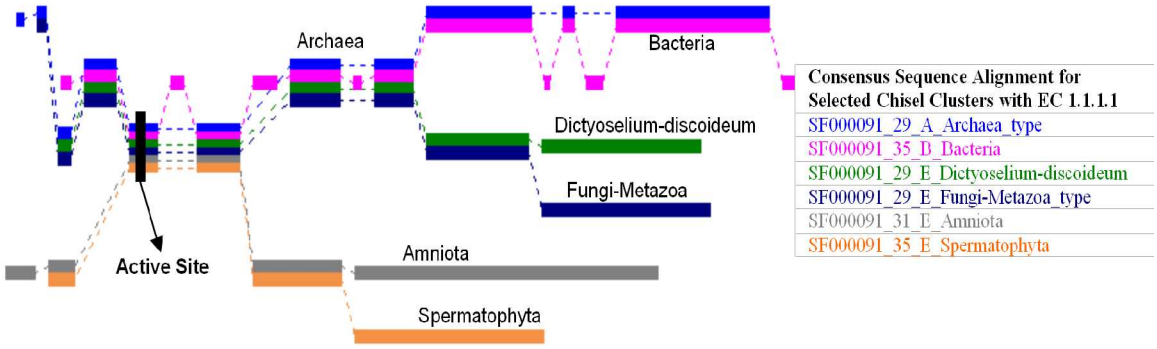


Fig 1. The POAVIZ alignment (Lee *et al*., 2002; Grasso *et al*., 2003) of the consensus sequences for the Chisel clusters derived from the superfamily of Zinc-containing alcohol dehydrogenases (PIR superfamily SF000091). Each color represents a consensus sequence for a cluster of enzymatic sequences corresponding to different taxonomc groups of organisms. The alignment demonstrates the conservation of the active site throughout the Chisel cluster consensus sequences; however, there is substantial variability in N- and C-terminus of the sequences depending on their taxonomic origin.

As it follows from the figure while sharing some conserved regions the sequences have undergone significant modifications to accommodate the needs of different taxonomic groups for adaptation.

The talk will describe the Chisel system and present the examples of its use for automated high-throughput genetic sequence analysis of 12 newly sequenced strains of Shewanella, interpretation and "binning" of the metagenomic data and for the needs of biomedical application (e.g. selection of antimicrobial drug targets, microbial diagnostics).

References:

Grasso, C. *et al*. (2003) POAVIZ: A partial order multiple sequence alignment visualizer. *Bioinformatics*, 19, 1446-1448.

Lee, C. *et al*. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18, 452–464.

Zhang, C.T. and Chou, K.C. (1995) An analysis of protein folding type prediction by seed-propagated sampling and jackknife test. *J. Protein Chem.*, 14, 583–59.

From cDNA to integrative protein annotation and beyond:

application to *Alvinella pompejana* cDNA collection

Gagnière N* [1], Bigot Y [2], Gaill F [3], Higuet D [4], Jollivet D [5], Leize E [6], Rees JF [7], Weissenbach J [8], Zal F [9], Poch O [1], Lecompte O [1]

1. Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, (CNRS/INSERM/ULP), BP 163, 67404 Illkirch Cedex, France.
2. Laboratoire d'Etudes des Parasites Génétiques, Université François Rabelais, Parc de Grandmont, 37200 TOURS, France.
3. Systématique, Adaptation, Evolution - Adaptation en Milieux Extrêmes, Université Pierre & Marie Curie, 75005 Paris, France.
4. Systématique, Adaptation, Evolution - Génétique et Evolution, Université Pierre & Marie Curie, 75005 Paris, France.
5. Equipe Evolution et Génétique des Populations marines, Station Biologique de Roscoff, Place Georges Teissier, BP74, 29682 Roscoff Cedex, France.
6. Laboratoire de Spectrométrie de masse BioOrganique, ECPM, 25 rue Becquerel, 67087, Strasbourg, France
7. Laboratoire de Biologie Cellulaire, Institut des Sciences de la vie - Université catholique de Louvain, 1, Place de l'Université B-1348, Louvain-la-Neuve, Belgique.
8. Génoscope, 2, rue Gaston Crémieux, CP 5706, F-91057 Évry Cedex, France.
9. Equipe Ecophysiologie : Adaptation et Evolutions Moléculaires, Station Biologique de Roscoff, Place Georges Teissier, BP74, 29682 Roscoff Cedex, France.

*To whom correspondence should be addressed: gagniere@igbmc.u-strasbg.fr

1. INTRODUCTION

Thanks to current high-throughput sequencing technologies, we now obtain a considerable amount of reliable sequence data, including complete genome of eukaryotic model organisms. In parallel, there is a multiplication of cDNA sequencing projects. These sequencing projects constitute an irreplaceable source of information about gene expression (tissue, development stage…) and allow bypassing problems of gene prediction. Moreover, they allow the exploration of the diversity of the tree of life at an acceptable cost by providing a gene overview for non-model organisms of interest. However, cDNA projects also induce specific problems such as errors due to single pass sequencing, incomplete cDNA sequences, difficulty to discriminate paralogs, splicing variants and polymorphism. This is especially true for “exotic” organisms exhibiting few homologous sequences in public databases.

Alvinella pompejana, the most thermotolerant metazoan known, is one of these singular organisms. This deep-sea worm lives in an extreme and variable environment in terms of temperature, pH and oxidation. It belongs to Annelids phylum – the segmented worms – for which very few sequences are available. To obtain phylogenetic and adaptative data as well as a pool of thermotolerant proteins for structural studies, a massive cDNA sequencing project has been initiated at the Centre National de Séquençage (Genoscope - Evry, France). To exploit the sequences, we have designed a dedicated relational database and developed a semi-automated protocol starting from *Alvinella* cDNA collection up to annotated proteins. This protocol includes chromatograms base-calling, raw sequences cleaning and assembling as well as original strategies for protein creation and annotation.

More than 63,000 cDNAs from distinct *Alvinella* tissues have been processed with our assembly protocols leading to 3,118 contigs and 6,799 singulets. The annotation relies on hierarchized multiple alignment of complete sequences (MACS) issued from each query protein and provide information at protein family and subfamily levels. This innovative approach allows sequence verification, correction and validation as well as propagation of annotation. The first step is the mining of various information (taxonomic data, functional descriptions, structural domains, secondary structures, active site residues…) from public databases by MACSIMS (1), a MACS Information Management System. In addition, accurate sequence-based *ab initio* predictions – such as coils, low complexity and transmembrane domains – are performed. These different types of data are integrated in the framework of the multiple alignment, allowing reliable data validation and rational propagation of information from the known to the new sequences. MACSIMS logs all the results in a XML file format for easy parsing and integration into other protocol workflows. This annotation

by MACSIMS is completed by a data mining program that generates a consensus functional definition from close homologs. The protein is also classified according to the Gene Ontology by GOAnno (2), a MACS-based Gene Ontology annotation program. With this protocol, about 90% of 5,050 assembled sequences with homology have been annotated with either functional definition, PFAM-A domains or Gene Ontology.

The relational database has been specifically designed to maintain fine grained information about cDNAs assembly processes (sequence quality, cloning errors, biased composition, tissular origin…) as well as to facilitate tissue libraries comparison, variant comparison and efficient exploitation of the *A. pompejana* cDNAs. Notably, we have developed a new program, OliDA based on MACSIMS extracted knowledge to automatically determine the protein domain boundaries, the best cDNAs and to design the oligos for protein expression tests for crystallization experiments.

Besides the annotation, by maintaining consistency between all the data produced, this integrated protocol has proved to greatly facilitate and improve the automated analysis and exploitation of a cDNA sequencing projects.

2. REFERENCES

1. Thompson, J.D., Muller, A., Waterhouse, A., Procter, J., Barton, G.J., Plewniak, F. and Poch O. 2006. MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics* 23;7:318

2. Chalmel, F., Lardenois, A., Thompson, J.D., Muller, J., Sahel, J.A., Leveillard, T. and Poch, O. 2005. GOAnno: GO annotation based on multiple alignment. *Bioinformatics* 1;21(9):2095-6

Protein function prediction in *Arabidopsis thaliana* using Bayesian networks and the Gene Ontology

James R. Bradford[1], Chris J. Needham[2], Andy J. Bulpitt[2], David R. Westhead[1]*
[1]Institute of Molecular and Cellular Biology, University of Leeds, Leeds, LS2 9JT, UK
[2]School of Computing, University of Leeds, Leeds, LS2 9JT, UK
*To whom correspondence should be addressed: D.R.Westhead@leeds.ac.uk

1. INTRODUCTION

We apply a Bayesian network based methodology to the challenging problem of assigning functions (described by the Gene Ontology (GO) (1)) to proteins of *Arabidopsis thaliana*. Compared to *Saccharomyces cerevisiae*, on which previous methods have focused (2)(3), *Arabidopsis* is poorly annotated consisting of over 30000 unique gene products: 47% of these have an unknown molecular function, and 49% and 64% have yet to be assigned to a cellular component and biological process respectively. By integrating heterogeneous data sources such as sequence motif, protein-protein interaction (PPI) and gene expression data, and utilizing the Bayesian network's ability to handle these often noisy, uncertain and incomplete data, we can accurately assign GO terms in all three functional categories.

2. METHODS

Bayesian network structure. The GO category (molecular function, cellular component, biological process) is represented by a sub-graph of GO terms, which forms a directed acyclic graph (DAG) to be used within our Bayesian network classification scheme. Only those GO terms with a gene association count above a certain threshold are included in each sub-graph. Relationships between nodes are captured by the directed edges of the Bayesian network. Edges between GO term nodes and their corresponding features are included as well as edges between parent and child GO terms nodes where parent nodes represent more general functional descriptions.

Feature representation. The set of GO annotations and features for each protein are described by a set of binary variables (one for each node in the graph). A node's variable is set to one if the GO term/feature it represents is included in the description of that protein, and zero if not. Currently, features can take the form of Interpro entries, interacting proteins or co-expressed genes.

Motif data. 26751 protein sequences based on the TAIR6 (4) release of the *Arabidopsis* genome were queried against the Interpro database (5) using Interproscan (6). Output consisted of a list of Interpro entries corresponding to each protein sequence.

Gene expression. Data from over 2400 microarrays were collected from the European *Arabidopsis* Stock Centre (7). Co-expression was measured by calculating the Pearson correlation coefficient (PCC) between all possible pairs of ~20000 genes represented on each microarray. Gene pairs with PCCs above a suitable threshold were deemed "co-expressed".

Protein-protein interactions. We are currently assessing the usefulness of a dataset of ~24000 potential protein-protein interactions in *Arabidopsis* (downloaded from TAIR (4)). This dataset is entirely derived from ortholog matching by Inparanoid (8) using yeast, fly, worm and human interactome datasets.

Learning and evaluation. The Bayesian network is trained on data consisting of a set of gene products with known function. Associated with each gene product is their GO term membership (whether the gene product is annotated with a particular GO term or not), and feature membership (whether a feature is present or absent in the gene product). Once the model parameters are learned, the likelihood of an annotation for a query gene product may be calculated. To do this, the joint probability of each possible most specific single GO annotation and the query features is computed. Figure 1 visualises these probabilities on the GO graph.

3. FIGURES

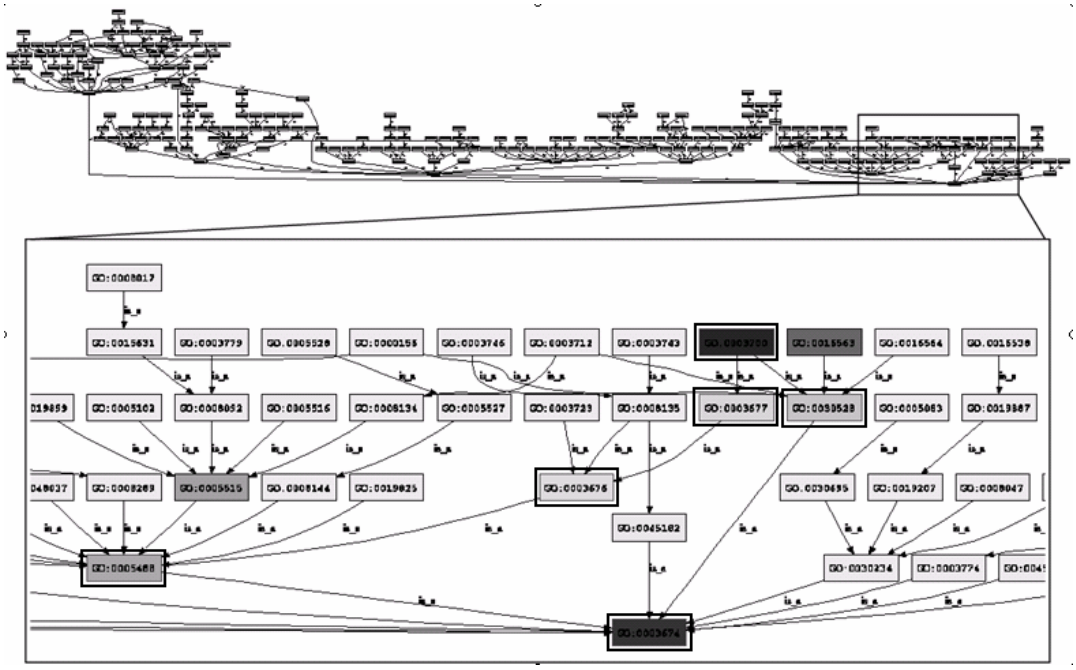


Figure 1: Protein function prediction onto the Gene Ontology. Top: 268 molecular function GO terms for which there are more than 20 gene associations in *Arabidopsis*. In these graphics more general terms are lower, with the root molecular function node at the base. Bottom: Zoomed in on the section of the GO graph containing predictions (using sequence motif data only) with non- or close-to-zero probability for a query gene. The double-bordered boxes denote the ground truth annotation for the query gene (including parent terms), and the shading denotes the probability of our predictions of the term being the most specific annotation for that gene; the darker the shading, the higher the probability.

4. RESULTS AND DISCUSSION

The training and testing of this Bayesian network schema on motif data only has produced encouraging early results. Furthermore, preliminary analysis has also found that the heterogeneous data types complement each other in terms of performance on each GO category. For example, gene expression data is particularly predictive of cellular component and biological process, whilst motif data is more predictive of molecular function. Future work will involve applying the trained Bayesian network to proteins of unknown function in *Arabidopsis*, and collaborating with experimentalists in order to verify interesting predictions.

5. REFERENCES

1. The Gene Ontology Consortium 2000. Gene Ontology: tool for the unification of biology. *Nature Genet.* 25: 25-29

2. Troyanskaya, O.G. et al. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc. Natl Acad. Sci. USA, 100, 8348–8353.

3. Nariai, N., Kolaczyk, E.D., Kasif S. 2007 Probabilistic protein function prediction from heterogeneous genome-wide data. PLoS ONE 2: e337

4. The *Arabidopsis* Information Resource, http://www.arabidopsis.org.

5. Mulder, N.J. et al. 2007. New developments in the InterPro database. *Nucleic Acids Res.* 35: D224-D228.

6. Quevillon, E., Silventoinen, V.., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res.* 33: W116-W120.

7. The European Arabidopsis Stock Centre, http://arabidopsis.info/.

8. Remm, M., Storm, C.E.V. & Sonnhammer, E.L.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, 314:1041-1052.

Protein function annotation in UniProtKB/Swiss-Prot and its use for automated function prediction

Brigitte Boeckmann[1]*, Lydie Bougueleret[1] and Amos Bairoch[1,2]

[1] Swiss Institute of Bioinformatics, 1 rue Michel Servet, 1211 Geneva 4, Switzerland
[2] Department of Structural Biology and Bioinformatics, University of Geneva, Switzerland
* To whom correspondence should be addressed

1. INTRODUCTION

Many procedures for automated function prediction perform a sequence similarity search and subsequently extract function-related annotation from relevant database entries. This approach requires a good understanding of content of databases: Does the protein function indicated in a database entry actually correspond to the given protein sequence? Where and what type of functional information can be found? Which information originates from experimental results, which ones from automated or automatic annotation? Does a database perform function prediction in a conservative manner in order to increase the reliability of the database content? Or does it maximize the information content, thus accepting a higher number of false positive annotations?

A widely used database for information retrieval is the UniProtKB/Swiss-Prot Protein Knowledgebase [1,2]. The questions raised above are discussed below.

2. PROTEIN VARIETY AND FUNCTIONAL DIVERSITY

Automated function prediction procedures generally assign protein function to amino acid sequences that correspond to a translated coding sequence. However, things are not so simple. A great majority of amino acid sequences need to be chemically altered in multiple ways and/or to assemble with other proteins before performing a function which none of the components of the complex could perform on its own [3]. Furthermore, cellular systems have developed multiple ways of modulating a protein's underlying genetic information – such as by alternative promoter usage or alternative splicing – and we are only beginning to anticipate the amplitude of the consequences. This implies a general need for a further refinement of automated annotation procedures and a more detailed structure for sequence databases.

Moreover, protein function can be assigned at distinct levels, e.g. molecular function, cellular role and phenotypic expression. All these functional classes are annotated in UniProtKB/Swiss-Prot, where the GO ontology [4] refers specifically to a protein's molecular function.

3. ASSIGNMENT OF FUNCTIONAL ANNOTATION IN UniProtKB/Swiss-Prot

All knowledge on protein function derives from experimental studies and it is mostly obtained from published results. In addition to the literature input, a variety of prediction tools are applied to every protein sequence that enters the knowledgebase, and experimental results and predictions are checked for their coherence when manually annotated. Annotation which is only based on predictions is tagged by non-experimental qualifiers.

The automated annotation procedure (Anabelle [5]) corresponds to manual annotation, with the only difference that annotation blocks linked to predictions are linked to rules, which are identical to those used for manual annotation; this includes, for example, the verification of biologically relevant positions within domains. Annotation rules often include further refinement of the analysis and ensure standardized annotation through all the taxonomic-based annotation projects.

HAMAP is a system that identifies and semi-automatically annotates proteins that are part of well-conserved families or subfamilies [6]. For each HAMAP family, initial members are chosen manually to produce a 'seed alignment' for the automatic generation of a profile. This profile is used to detect further possible microbial family members in UniProtKB. Function-related information retrieved from the literature is used to create a HAMAP family rule manually. HAMAP rules are then applied to bacterial, archaeal and plastid-encoded proteins. There are many filters which distinguish between typical family members that can be annotated automatically and those which need manual checking.

Currently, automatically annotated entries are still manually checked before their integration into UniProtKB/Swiss-Prot. Both systems, Anabelle and HAMAP, use common modules for automated analysis, the selection of analysis results and the creation of annotation.

4. EXTRACTION OF FUNCTIONAL INFORMATION FROM UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot entries contain protein function annotation in various forms and at distinct levels. Although protein names in the DE (DEscription) line frequently hint on a protein's function since it also includes the EC number, ontologies are a more appropriate source. Swiss-Prot includes two ontologies: keywords and Gene Ontology (GO) annotation. Keywords are controlled vocabularies; function-related keywords indicate a protein's function at low resolution. In contrast, GO terms can be very specific and the hierarchical structure of the GO ontology makes it possible to assign a function at a trusted level of specificity. The GO annotation in UniProtKB/Swiss-Prot contains evidence tags, which are valuable to assess the reliability of the annotation obtained. However, GO terms are not linked to protein isoforms. An enzyme's molecular function is described under the CC line topic CATALYTIC ACTIVITY, and controlled vocabularies indicate the pathway context under the CC line topic PATHWAY. A more descriptive function-related annotation is also available in the CC line topics FUNCTION. Further functional annotation available in UniProtKB/Swiss-Prot will be presented.

5. OUTLOOK

Relevant recent and upcoming format modifications in UniProtKB/Swiss-Prot are presented, e.g. the new PE line-type used to indicate different experimental evidence relative to a protein's existence.

6. ACKNOWLEDGEMENTS

The authors would like to thank Vivienne Baillie Gerritsen for the correction of the manuscript. This work is partially funded by the Swiss Federal Government through the State Secretariat for Education and Research, the European Commission within the Research Infrastructure Action of the FP6 "Structuring the European Research Area" Programme, contract number 021902 (RII3) and the National Institutes of Health (NIH) grant 2 U01 HG002712-04.

7. REFERENCES

[1] Wu CH, Apweiler R, Bairoch A, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 2006;34:D187-D191.

[2] Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31:365-370.

[3] Boeckmann B, Blatter MC, Famiglietti L, et al. Protein variety and functional diversity: Swiss-Prot annotation in its biological context. C. R. Biol. 2005;328:882-899. Epub 2005 Jul 28.

[4] Harris MA, Clark J, Ireland A, et al. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004 Jan 1;32:D258-261.

[5] Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-Prot: juggling between evolution and stability. Brief Bioinform. 2004;5:39-55.

[6] Gattiker A, Michoud K, Rivoire C et al. Automated annotation of microbial proteomes in SWISS-PROT. Comput. Biol. Chem. 2003;27:49-58.

Dimensionality Reduction for Protein Function Prediction

Zehra Cataltepe[1,*], Eser Aygun[1], Asli Filiz[1], Ozlem Keskin[2], Caner Komurlu[1],Yucel Altunbasak[3]
[1]Istanbul Technical University, Computer Engineering Department
Ayazaga, Sariyer, TR-34469, Istanbul, Turkey
[2]Koc University, College of Engineering, Chemical and Biological Engineering
Rumeli Feneri Yolu, Sariyer, TR-34450, Istanbul, Turkey
[3]Georgia Institute of Technology, School of Electrical and Computer Engineering
Atlanta, GA 30332-0250
*To whom correspondence should be addressed: cataltepe@itu.edu.tr

1. INTRODUCTION

Dimensionality reduction methods (see [1], for example) reduce the input feature dimensionality and result in faster classification algorithms due to smaller number of inputs. If noisy, irrelevant or redundant features are eliminated, then dimensionality reduction could also lead to better classification accuracy.
Previously, [2] compared a number of feature selection methods, based on entropy, t-statistics and chi^2 statistics and found out that dimensionality reduction helped with distinguishing genes for Acute Lymphoblastic Leukemia and ovarian cancer. [3] used manual or pairwise Fisher's Linear Discriminant Analysis to select features for identifying marker genes on a number of cancer data sets. [4] and [5] used sequence data together with Support Vector Machines for both feature and instance selection for protein function prediction.
In this study, we evaluate the effects of dimensionality reduction on protein function prediction. We consider the GO (Gene Ontology) Molecular Function first level categories and H. Pylori as the organism to collect the amino acid sequence data. We evaluate three different classifiers (Naïve Bayes, kNN, SVC) and three different dimensionality reduction methods (PCA, Fisher's LDA and FCBF (Fast Correlation-Based Filter) algorithm [6]).

2. DATA AND METHODS

We downloaded the fasta sequences for H. Pylori from http://expasy.org/sprot/hamap/HELPY.html There were a total of 564 amino acid sequences. The distribution of these amino acid sequences according to the GO Molecular Function categories are shown in Table 1. We eliminated GO Molecular Function categories with less than 10 sequences.
For features, we obtain physiochemical properties of the amino acid sequences using the PROFEAT software [7]. We also use the ClustalW [8] alignment scores between a sequence and all the training sequences. The dimensionality reduction algorithms that we consider are, PCA (Principal Component Analysis), Fisher's LDA (Linear Discriminant Analysis) and FCBF (Fast Correlation-Based Filter) algorithm. We measure the 10-fold cross validation accuracy of Naive Bayes (NB), kNN (k-Nearest Neighbor) and Support Vector Classifiers (SVC) to compare accuracy of classifiers that use features selected by different feature selection methods.

3. RESULTS

Features selected by FCBF: In our experiments, out of the 1447 PROFEAT features, 5 were selected by FCBF (see Table 2). In Table 2, AURM and AURV are features related to the Normalized Moreau-Broto autocorrelation. Out of 564 ClustalW features (i.e. sequences), 14 were selected by FCBF.

Classification Accuracies: Table 3 shows the average 10-fold cross validation accuracies when H. Pylori and PROFEAT features and H. Pylori and ClustalW features are used respectively. As seen in the tables the best accuracies for a single classifier are achieved when FCBF is used for all four cases. FCBF seems to work especially well for the Naive Bayes classifier. Although its performance is also quite good for kNN and SVC classifiers also. The reason why FCBF works very well with NB classifier may be the fact that NB assumes that each feature is independent from each each other and after elimination of redundant features, the remaining features are actually not correlated with each other. We think that FCBF works better in general for each classifier, because it uses an entropy based measure for redundancy and relevance. PCA or Fisher's LDA, on the other hand, use linear correlations. Entropy based measure is able to capture more general correlations between features and data and in between features.

Classification Time: On the average, FCBF takes less time to execute than the PCA and Fisher's LDA.

4. CONCLUSIONS

While feature selection could result in drastic falls in accuracy for a certain classifier when PCA or Fisher's LDA are used, there is either a very good improvement or very little drop in accuracy when FCBF is used. Therefore, when classification accuracies are averaged over all three classifiers (NB, KNN and SVC), FCBF performs better. Furthermore, in all the experiments, FCBF (usually with Naive Bayes as the classifier) results in the best accuracy. FCBF is also faster than PCA or Fisher's LDA. We did not observe a significant difference in classifier accuracies when either ClustalW alignment scores or PROFEAT features were used.

In the future, we would like to soften the FCBF algorithm, select more features and see the overlap between features when different organisms are used. It would be interesting to see if function is determined by different or similar features for each organism.

5. REFERENCES

1. Yang, Y. and Pedersen, J. O. 1997. A Comparative Study on Feature Selection in Text Categorization, *Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420.

2. Liu, H., Li, J. and Wong, L. 2002. A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns, *Genome Informatics*, 13, 51-60.

3. Wang, J., Hellem, T., Jonassen, I. et. al. 2003. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data, *BMC Bioinformatics*, 4:60.

4. Al-Shahib, A., Breitling, R. and Gilbert, D. 2005. Feature Selection and the Class Imbalance Problem in Predicting Protein Function from Sequence, *Applied Bioinformatics*, 4:3, 195-203.

5. Al-Shahib, A., Breitling, R. and Gilbert, D. 2005. FrankSum: new feature selection method for protein function prediction, *Int J Neural Syst.*, 15:4, 259-275.

6. Yu, L. and Liu, H. 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research* , 5, 1205-1224.

7. Li, Z. R., Lin, H. H., Han, L. Y. et al. 2006. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, *Nucleic Acids Research*, 34, W32-W37.

8. Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, 22:22, 4673-4680.

TABLE1: No of Sequences in Each Class.

GOID	Molecular Function	#seq.
0003774	motor activity	12
0030528	transcr. reg. act.	15
0004871	signal transd. act.	23
0003674	molec. func. (others)	26
0005215	transporter activity	72
0005488	binding	280
0003824	catalytic activity	334

TABLE2. PROFEAT features selected by FCBF

PROFEAT ID	Feature name
F5.1.1.1	Composition of Hydrophobicity(1)
F6.1.2.26	Seq. Order Coupling Number2(26)
F5.1.6.3	Composi. of Secondary structure(3)
F1.2.1.43	DD Dipeptide composition(%)
F2.1.8.15	AURM(15)
F2.1.6.6	AURV(6)

TABLE3. Classification Accuracies Using PROFEAT and ClustalW Features.

	Naïve Bayes		kNN		SVC	
Features	PROFEAT	ClustalW	PROFEAT	ClustalW	PROFEAT	ClustalW
all features	3.40 ± 0.34	2.05 ± 0.20	46.07 ± 2.11	47.18 ± 1.82	45.13 ± 1.17	45.13 ± 1.17
FCBF	**52.75** ± 1.85	**50.95** ± 2.20	41.79 ± 1.98	44.81 ± 2.21	43.28 ± 2.55	45.13 ± 1.17
PCA	44.57 ± 1.54	35.87 ± 1.55	34.72 ± 2.28	25.77 ± 1.30	25.67 ± 5.19	45.13 ± 1.17
FISHER	9.62 ± 1.02	3.25 ± 0.83	31.92 ± 1.78	7.99 ± 1.63	45.13 ± 1.17	45.13 ± 1.17

PFFmot – An integrated resource for characterizing structural and functional regions in proteins

Manuel Corpas*, David Thorne, James Sinnott, Steve Pettifer and Terri Attwood
University of Manchester, Faculty of Life Sciences & School of Computer Science, UK
*To whom correspondence should be addressed: corpas@bioinf.man.ac.uk

1. INTRODUCTION

Despite decades of work, understanding how proteins fold and function remains a major research challenge. The fruits of this massive research effort have been: development of (i) methods for predicting the likely structures and functions of protein sequences, or for simulating the folding process; and (ii) databases of structural and functional information (e.g., containing 3D coordinates, fold classifications, functional sites, family relationships, and so on). As part of the ongoing endeavour to understand the principles of protein structure and function, we have been involved in the development of a new, integrated resource, based on independent analyses of a subset of proteins from the PDB [1]. The motivation for combining data from these different analytical approaches was to offer insights into the role of particular types of residues and fragments in protein folding and function, and hence to improve our understanding of factors that are critical to the folding process and hence to the elicitation of protein function in general.

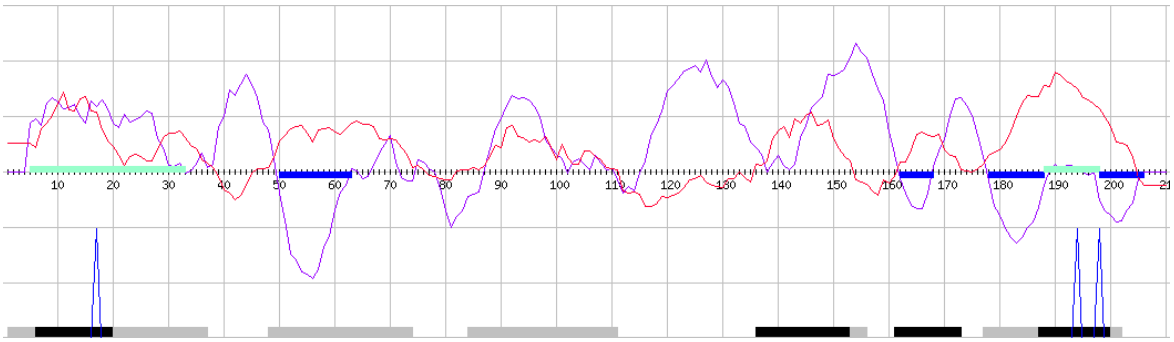
2. METHODS

The analysis methods chosen for study included residue accessibility values (residue water exposed surface in Angstrom^2), Fold-X [2] scores (denoting the stabilizing contributions to the fold), Popmusic [3] values (evaluates the changes in stability of a given protein or peptide under all possible single site mutations), lattice simulations [4] (denoting the number of close neighbors or interaction partners within the fold) and the degree of residue conservation. Each of these methods was applied to a dataset of 117 protein families (13,735 residues). All-against-all comparisons were made in order to identify those parameters that were most correlated, and a 'folding score' was created from a sum of normalised averages of the smoothed values. To avoid over-fitting the data, the folding score assumes that the relation between folding signals and conservation is linear. A similar conservation-based score was created combining conservation-related parameters Scorecons [5] and Crescendo [6]. The utility of the folding score in identifying structurally or functionally important regions was analysed by comparing the locations of known structural and functional motifs in the target proteins with the locations of folding score peaks and troughs.

3. RESULTS

An annotated database was generated using data from several unrelated algorithms and data sources, allowing an integrated analysis of structural and functional features. From an initial analysis of the data, we found, not surprisingly, that certain results were strongly correlated: *e.g.*, residue accessibility values, Fold-X scores, Popmusic values, and lattice simulations. The 'folding score' synthesized from these values was used to identify structurally and functionally important regions within protein sequences. Structurally favorable regions, as denoted by the folding score, tended to be well conserved; conversely, well-conserved but structurally unfavorable regions tended to be associated with functional sites. These observations suggested that we might be able to use the derived folding score automatically to pinpoint likely structural or functional regions.

4. FIGURE 1



Chloramphenicol Acetyltransferase Type III (PDB code: 3cla). The folding score is shown in purple across the 2D representation of the protein. In regions of high conservation (red profile), folding score troughs that (blue bars) pinpoint regions that are structurally favorable, while folding score peaks (green bars) indicate potential functional regions. Conservation scores were calculated from the alignment between the PDB sequence and representative family members from Swiss-Prot. Interestingly, the known catalytic residues (R18, H195 and D199), shown as blue spikes, fall within folding score peaks, lending weight to the notion that conserved, structurally destabilizing regions are likely to be functionally significant. At the foot of the diagram, manually-selected motifs (black and grey bars) are shown for comparison (the different colors reflect the fact that different motifs are representative of families and subfamilies with the protein family hierarchy): we see a tantalizing correspondence between the locations of likely structural and functional regions (blue and green bars) and known motifs (black and grey bars).

5. CONCLUSION

Using a combination of algorithms, we have created a consensus tool that can identify likely structural and functional regions in protein sequences. The integration of different methods adds value over individual approaches, and offers a means of automatic motif detection. This approach can therefore facilitate motif-based methods of protein family characterisation, allowing more informed decisions regarding motif selection.

5. AVAILABILITY

Version 1.0 of the PFF dataset is accessible in a DSSP-flat-file format from http://babylone.ulb.ac.be/LIFE/; it is also available in an XML format through the UTOPIA toolkit, together with the UTOPIA visualisation tools for OS X, Windows and Linux at http://utopia.cs.manchester.ac.uk. The Web resource for calculating combined folding scores is accessible at http://umber.sbs.man.ac.uk/corpas/pff/

6. REFERENCES

1. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
2. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W382-8.
3. Gilis, D., et al., *In vitro and in silico design of alpha1-antitrypsin mutants with different conformational stabilities.* J Mol Biol, 2003. **325**(3): p. 581-9.
4. Papandreou, N., et al., *Universal positions in globular proteins.* Eur J Biochem, 2004. **271**(23-24): p. 4762-8.
5. Valdar, W.S., *Scoring residue conservation.* Proteins, 2002. **48**(2): p. 227-41.
6. Chelliah, V., et al., *Distinguishing structural and functional restraints in evolution in order to identify interaction sites.* J Mol Biol, 2004. **342**(5): p. 1487-504.

Automatic classification of protein phylogenetic profiles

Juan Carlos Fernandez [1,2]*, Jesus K. Estrada-Gil [1,2],
Enrique Hernandez-Lemus [2], Enrique Morett [4], Gerardo Jiménez-Sánchez [3], Edgar E. Vallejo[1]
[1] Computer Science Department, ITESM Campus Estado de Mexico, Carretera Lago de Guadalupe km 3.5, Atizapan de Zaragoza, 52926, Mexico.
[2] Department of Computational Genomics and [3] Department of Basic Research, Instituto Nacional de Medicina Genómica, Periferico Sur No. 4124, Torre Zafiro 2, Piso 5,D.F., 01900, Mexico.
[4] Department of Cellular Engineering and Biocatalysis, IBT UNAM, Av. Universidad 2001, Cuernavaca, 62210, Mexico.
*To whom correspondence should be addressed: jfernandez@inmegen.gob.mx

1. INTRODUCTION

Computational intelligence techniques have been extensively employed together with post-genomic approaches for the automated prediction of functional coupling of proteins (1). One of the later methods, called protein phylogenetic profiles (PPP), describe the correlated presence and absence of genes among a set of organisms and has proven to be particularly effective for detecting a variety of functional associations in the protein space (2). This method relies crucially on the clustering method employed for reconstructing functionally related groups of proteins. Its principal weakness is that proteins are often associated with unrelated groups that possess very similar phylogenetic profiles (i.e. false positive problem). Previous approaches have been developed to increase the accuracy of this method (3-6). However, the main limitation of the clustering algorithms used so far has been the lack of *expressiveness* for characterizing the participation of proteins in more than one functional pathway.

We have previously developed an approach to improve the expressiveness of the PPP method (7). In this work, we extend our approach with an automatic classification method useful to infer functional coupling of new and/or hypothetical proteins. We first used the fuzzy C-means clustering algorithm (FCM) (7, 8) that represents a convenient framework for describing associations between proteins. FCM constructs a fuzzy partition of a data set that provides a way in which the strength of association of grouped elements to a particular cluster is given. This association is described by a continuous value between 0 and 1 that provides information on the membership of each PPP included in the data set to each cluster. These membership values are then used to train neural network for the classification of PPPs.

We constructed a training set of 2,223 phylogenetic profiles obtained from the Cluster of Orthologous Groups of Proteins database (11) grouped by the FCM algorithm (7). This information was used to choose the 1,040 most informative phylogenetic profiles, after conducting extensive experiments on clustering phylogenetic profiles with FCM optimized using Shannon's Information Theory (9). The fuzzy clustering produced was applied to train an artificial neural network with a radial basis function (ANN-RBF) (10). We evaluated the predictive power of our approach to infer membership values to a given phylogenetic profiles cluster in a 10-item validation set not included in the 1030-item (within a membership matrix approach) training set. Table I and II show the automated assignation of the phylogenetic profiles with low difference to the membership value expected, including four phylogenetic profiles correctly classified in the first group. These results suggest that ANN-RBF is able to detect the hypothetically new protein phylogenetic profiles even if they were not included at the initial clustering experiments.

Preliminary results indicated that FCM is capable of detecting non-trivial relationships among proteins. In general, we believe that fuzzy set theory and artificial neural networks provide an appropriated framework for characterizing biological processes. As future work, we expect to include a larger Cluster of Orthologous Groups of Proteins database (12) from a single domain to increase the clustering accuracy and use the Support Vector Machine (SVM) method (13) to infer hypothetically new phylogenetic profiles. This studies hold the promise to contribute to elucidate the fundamental mechanisms of protein evolution.

TABLE I VALIDATION
Original phylogenetic profiles associations (groups and memberships values)

Phylogenetic profile	COG	G1	Value-1	G2	Value-2	G3	Value-3	G4	Value-4	G5	Value-5
11111111001010011101110000	**COG1793**	**3**	0.01830	**11**	0.01716	**61**	0.01642	**5**	0.01641	**1**	0.01640
11101001010111000000000111	**COG1624**	**9**	0.01783	**20**	0.01702	**35**	0.01639	**14**	0.01636	**67**	0.01584
10101000110010011100010100	**COG1392**	**12**	0.01781	**50**	0.01699	**67**	0.01693	**30**	0.01689	**34**	0.01618
00011010111111111010110000	**COG0726**	**21**	0.01778	**46**	0.01776	**1**	0.01686	**70**	0.01682	**68**	0.01680
01101011011111111110111010	**COG0671**	**25**	0.01759	**64**	0.01759	**33**	0.01754	**1**	0.01655	**39**	0.01655
00100011010111111101110001	**COG1087**	**33**	0.01829	**2**	0.01728	**68**	0.01719	**7**	0.01650	**48**	0.01647
11010100111111111100010000	**COG1441**	**42**	0.01749	**28**	0.01745	**46**	0.01742	**24**	0.01646	**3**	0.01642
01000001111010111100110000	**COG1414**	**48**	0.01694	**3**	0.01605	**2**	0.01605	**31**	0.01538	**64**	0.01538
00011110000110000100100000	**COG0819**	**54**	0.01732	**57**	0.01730	**9**	0.01667	**51**	0.01662	**34**	0.01661
11111111110111011000110000	**COG0075**	**70**	0.01912	**46**	0.01773	**38**	0.01767	**11**	0.01765	**1**	0.01671

TABLE II ANN-RBF
Simulation results

Phylogenetic profile	COG	G1	Value-1	G2	Value-2	G3	Value-3	G4	Value-4	G5	Value-5
11111111001010011101110000	**COG1793**	**61**	0.01574	**5**	0.01571	**69**	0.01561	**24**	0.01553	**11**	0.01544
11101001010111000000000111	**COG1624**	**36**	0.01638	**35**	0.01625	**47**	0.01511	**9**	0.01510	**54**	0.01500
10101000110010011100010100	**COG1392**	**36**	0.01661	**12**	0.01625	**30**	0.01602	**50**	0.01598	**47**	0.01535
00011010111111111010110000	**COG0726**	**46**	0.01822	**21**	0.01785	**1**	0.01633	**70**	0.01613	**31**	0.01539
01101011011111111110111010	**COG0671**	**64**	0.01740	**62**	0.01595	**10**	0.01577	**22**	0.01573	**23**	0.01560
00100011010111111101110001	**COG1087**	**33**	0.01765	**68**	0.01637	**21**	0.01556	**36**	0.01549	**23**	0.01536
11010100111111111100010000	**COG1441**	**42**	0.01772	**28**	0.01647	**17**	0.01629	**3**	0.01620	**46**	0.01580
01000001111010111100110000	**COG1414**	**48**	0.01578	**2**	0.01547	**36**	0.01543	**31**	0.01515	**30**	0.01506
00011110000110000100100000	**COG0819**	**36**	0.01630	**35**	0.01559	**54**	0.01556	**47**	0.01517	**57**	0.01505
11111111110111011000110000	**COG0075**	**70**	0.02027	**38**	0.01748	**46**	0.01717	**3**	0.01640	**36**	0.01554

2. REFERENCES

1. Eisenberg, D., Marcotte, et al. E. 2000 Protein function in the post-genomic era. *Nature*, Vol. 405, pp. 823-826.
2. Pellegrini, M. et al. 1999 Assigning protein function by comparative genome analysis: Protein phylogenetic profiles, *PNAS*, Vol. 96, pp.4285-4288.
3. Wu, J. et al. 2003 Identification of functional links between genes using phylogenetic profiles. Bioinformatics, Vol.19(2), pp.1524-1530.
4. Marcotte, E., Pellegrini, M. et al. 1999 A combined algorithm for genomewide prediction of protein function. *Nature*, Vol. 402, pp.83-86.
5. Marcotte, E. 2000 Computational genetics: finding protein function by nonhomology methods. *Current Option in Structural Biology*. Vol.10, pp.359-365.
6. Sali, A. 1999 Functional links between proteins. *Nature*, Vol.402, pp.23-26.
7. Fernández, J.C. et al, 2006, Fuzzy C-means for inferring functional coupling of proteins from their phylogenetic profiles, *IEEE CIBCB*, 1, 23-30.
8. Bezdek, J.C. 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press.
9. Shannon, C.E. 1948 The mathematical theory of communication, *Bell Syst, Techn. Journ.*, 27, 379-423; 623-656.
10. Jang J. et al. 1997 *Neuro-fuzzy and soft-computing*. Prentice Hall.
11. Tatusov, L. et al. 2001 The COG database: new developments in phylogenetic classification of protein from complete genomes. *Nucletic Acids Research*, Vol. 29, No.1.
12. Tatusov, R. L. et al, 2003 The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, Vol. 4, pp. 41-54.
13. Vapnik, V. 1979 Estimation of Dependences Based on Empirical Data. Nauka. In Russian. English Translation by *Springer Verlag* in 1982.

A Novel Conservation Score For Functional Residue Prediction

Julia D. Fischer[1], Andrei N. Lupas[1], Johannes Söding[1,2]
[1]Max-Planck Institute for Developmental Biology, Tübingen, Germany
[2]Gene Center, LMU Munich, Munich, Germany
{johannes.soeding, julia.fischer}@tuebingen.mpg.de

Abstract: Scoring positional conservation is a widely applied and powerful method in protein sequence analysis because residues that are important for a protein's function are likely to be conserved in homologous protein sequences. Despite the importance and many different propositions, no conservation score is known that would satisfy a set of desirable criteria [Val02]: (1) The score should be in the range [0,1]. (2) It equals 0 iff the amino acids are distributed like the background distribution. (3) It is maximum iff the column is totally conserved. (4) The score should take amino acid similarities into account. (5) It should penalize gap-rich columns.

We have developed a simple conservation score (FRcons) which fulfills the above criteria. Its usefulness is demonstrated in a large benchmark to predict ligand-binding residues in proteins (Fig 1). FRcons outperforms entropy-based, sum-of-pair-based and variance-based methods [PG01]. A further improvement is achieved by weighting down hydrophobic amino acids which are often conserved by structural rather than functional constraints. We will also present results on our next step, a functional residue predictor which integrates FRcons with solvent accessibility and secondary structure prediction in a Bayesian framework.

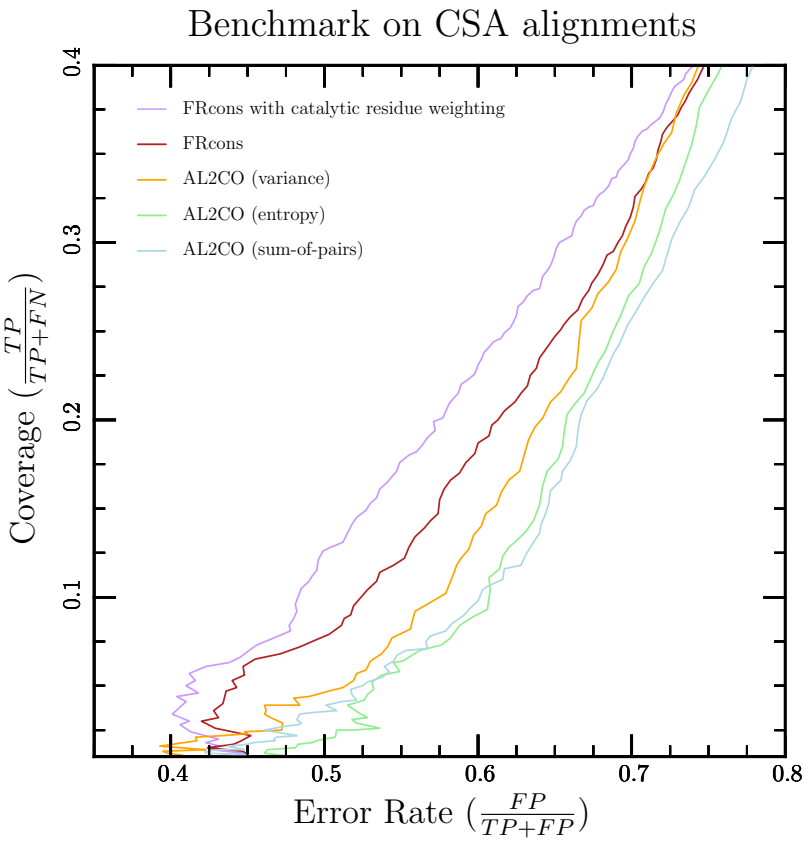


Figure 1: Benchmark results based on 416 proteins from the Catalytic Site Atlas (CSA [PBT04]). Positives are residues that are marked as catalytically active in the CSA and additionally all residues in a $4\mathring{A}$ radius of the ligand in the catalytic site. Every query sequence comprises only one SCOP domain. False positives are abbreviated FP, true positives TP and false negatives FN.

FRcons is available at our web server for functional residue prediction (http://frpred.tuebingen.mpg.de). The input can be a single sequence or an alignment. The server will mark the highest-scoring columns the alignement by color saturation. If a structure is available, a JMol applet shows the structure with conserved residues highlighted.

References

[PBT04] CT Porter, GJ Bartlett, and JM Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, 32(Database issue):D129–D133, 2004.

[PG01] J Pei and NV Grishin. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, 17(8):700–712, 2001.

[Val02] William S J Valdar. Scoring residue conservation. *Proteins*, 48(2):227–241, 2002.

A simplified representation of electrostatic model surfaces for addressing protein-protein interaction problems

Dietlind L Gerloff[1*], Shakir Ali[2], Xueping Quan[2], Rupert A. Koenig[3]

[1]: Dept. of Biomolecular Engineering, UC Santa Cruz; [2]: School of Informatics, University of Edinburgh;
[3]: University of Applied Sciences Bingen;
[*] to whom correspondence should be addressed: gerloff@soe.ucsc.edu

1. INTRODUCTION

Intense efforts in structural biology and genomics are rapidly producing template structures that allow structural bioinformaticians to model the 3-D structures of many more template proteins comparatively, "by homology". However, modelling does often not immediately reveal the molecular function of the target, especially in cases where the molecular function of the template protein is also unknown. Function prediction is particularly difficult if the molecular interactions did not give rise to amino acid conservation across at least subfamilies of homologues – unfortunately, this is not an uncommon situation when functions involve transient interactions. The electrostatic surface properties of protein structures can sometimes provide clues about molecular interactions in which they engage. Structural biologists often inspect the surface properties of a newly determined protein structure visually e.g. using the program GRASP (1), to bolster their confidence in hypotheses they may have regarding its interaction sites, or partners. It should prove particularly interesting in this context to compare the electrostatic surfaces of sets of homologous proteins (2, 3). Within such sets, all structures can be modelled with reasonable confidence if the 3-D structure of at least one protein is known and a reliable sequence alignment between each homologue and this protein can be generated. Even if we take into consideration that the surface properties in a comparative model can merely be an approximation of those in the true structure, it would be of great interest in the context of various protein-protein interaction problems to be able to compare electrostatic model surfaces automatically, i.e. systematically and effectively on substantive data sets. Only very few methods exist at present that allow such "screening-style" comparisons (2, 4).

2. RESULTS AND DISCUSSION

We have developed a novel way of simplifying the comparison of the electrostatic molecular surfaces of proteins to the comparison of **1-D "electrostatic surface profiles"** based on the same information. In these surface profiles the electrostatic surface charge of each protein is basically apportioned to its individual residues. At the conference I will present this method, and show two examples of how comparisons of the electrostatic model surfaces of homologous proteins using surface profiles may be useful in protein-protein interaction questions such as binding site and partner prediction (see below). The simplified 1-D representation that is introduced here will necessarily mean to "neglect" more fine-grained information. However, the profile format offers many advantages over the classic 3-D format of electrostatic potential surfaces. Most obviously, studies of the kind discussed here could be combined much more easily with multiple sequence analysis approaches of various kinds, for example correlated mutation analysis between potential partner proteins in protein-protein interactions.

Binding site prediction - complement receptor 1 (CR1):

Our results indicate that systematic comparisons of surface profiles are helpful for **pinpointing functionally important domains within a set** of homologous domains in the same protein (e.g. in the human immune-regulating protein complement receptor 1). CR1 is known to be involved in protein-protein interactions with several partners at several sites along its length (1998 amino acids). However, not all partners are known and the location of the binding sites on different domains, with respect to their common structural scaffold, can differ (5). Comparing the surface profiles of the models of the 30 homologous domains in CR1 (6) to each other, by reference to their sequence similarity, suggests which domain surfaces seem to have changed more than would be expected – which may reflect the acquisition of new interaction partners during evolution. Some details regarding how best to select the most

"outstanding" domains remain to be worked out better before this screening approach can be fully automated and generalised. However, already at this stage our **results agree well with visual inspection** of GRASP (1) pictures and can be compared to those obtained by other methods for electrostatic surface comparison. While experimental information about interactions between CR1 and other proteins is scarce, the domains pinpointed by our comparisons seem to be involved in such interactions. Indeed our results are **compatible with the current biological knowledge** regarding binding-site location, which is largely based on site-directed mutagenesis experiments.

Partner prediction - CDK-cyclin homologs in Arabidopsis thaliana:

Where families of paralogous proteins exist (i.e. homologs within the same species), not every member of the one protein family will necessarily interact with every member of the other protein family. In a previous study, we investigated the potential of a molecular docking approach with modelled protein structures for answering the question which are the **most likely interacting partners** in CDK-cyclin like complexes between the approximately 35 CDK and 50 cyclin homologs in *Arabidopsis thaliana* (6). In contrast to the interaction problem described above, the three-dimensional orientation of the two partner proteins in putative complexes can be assumed to be similar in all complexes formed in this case. However, it is impossible to associate each *Arabidopsis thaliana* CDK and cyclin confidently with any one of the well-studied human CDK and cyclin types based on their sequences. Therefore, the most plausible CDK and cyclin homolog combinations cannot be predicted by inference, i.e. based on known human CDK-cyclin pair interactions. One of the **outcomes of control studies** during this project was that **electrostatic complementarity** at the interaction interface between the partners can help tremendously to distinguish between likely interacting, and not interacting, complexes – in spite of these being transient complexes, which cannot be expected to be very stable *a priori*. Intersecting the results of molecular docking with electrostatic complementarity analysis using the program MOLSURFER (8) suggested 19 most likely interacting CDK-cyclin pairs out of 1188 possible pairs. An alternative prediction method for this problem using electrostatic surface profiles was derived, in which the profiles of all possible CDK-cyclin combinations were compared over the range of their interacting residues and examined for complementarity. While there is hardly any biological laboratory data against which we can validate the results by both approaches, their predictions can be compared to one another.

3. REFERENCES

1. Nicholls, A., Sharp, K. A. and Honig, B. 1991. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* 11: 281-296

2. Blomberg, N., Gabadoulline, R. R., Nilges, M. and Wade, R. C. 1999. Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. *Proteins* 37:379-387.

3. Pawlowski, P. and Godzik, A. 2001. Surface map comparison: studying function diversity of homologous proteins. *J. Mol. Biol.* 309:793-807.

4. Sasin, J. M., Godzik, A. and Bujnicki, J. M. 2007. SURF'S UP! - Protein classification by surface comparisons. *J. Biosci.* 32:97-100.

5. Kirkitadze, M. D. and Barlow, P. N. 2001. *Immunol. Rev.* 180: 146-161.

6. Soares, D. C., Gerloff, D. L., Syme, N. R., Coulson, A. F. W., Parkinson, J. and Barlow, P. N. 2005. Large-scale modelling as a route to multiple surface comparisons of the CCP module family. *Prot. Eng. Des. Sel.* 18: 379-88

7. Quan, X., Doerner, P. and Gerloff, D. L. Partner prediction for transient protein complexes: CDK-cyclin homologue interactions. *Manuscript submitted (2007).*

8. Gabdoulline, R. R., Wade, R. C. and Walther, D. 2003. MolSurfer: A macromolecular interface navigator. *Nucl. Acids Res.* 31:3349-3351.

Convergent evolution of enzyme active sites is not a rare phenomenon

Pier Federico Gherardini[1,2], Mark N Wass[1], Manuela Helmer-Citterich[2], Michael JE Sternberg[1*]
1 Biochemistry Building, Division of Molecular Biosciences, Imperial College
London, London SW7 2AZ, UK
2 Centre for Molecular Bioinformatics, Department of Biology,
University of Tor Vergata, Rome, Italy
*To whom correspondence should be addressed: m.sternberg@imperial.ac.uk

1. INTRODUCTION

The aim of this work is to systematically identify and characterise examples of convergent evolution of enzyme active sites and assess how common this phenomenon is across enzyme space. Two modes of convergent evolution are considered: mechanistic analogues are enzymes that use the same mechanism to perform related, but possibly different, reactions, while transformational analogues catalyse exactly the same reaction, but may use different mechanisms.

This work integrates detailed knowledge about catalytic residues, available through the recently developed Catalytic Site Atlas (CSA) (1), with the evolutionary information provided by the Structural Classification of Proteins (SCOP) database (2). Mechanistic analogues have been identified using the Query3d local structural comparison algorithm (3) to identify structural similarities in unrelated active sites. The results of this large scale comparison have subsequently been validated and analysed via an extensive search of the relevant literature. Transformational analogues were defined as those enzymes having the same EC number but belonging to different SCOP superfamilies.

2. RESULTS

The results of this analysis provide a hand-curated set of instances of convergent evolution gathered for the first time with a systematic approach. Different levels of mechanistic convergence are considered, ranging from identical mechanisms, as in the case of serine proteases, to enzymes that implement analogous chemical "strategies" in different ways. We used such a broad definition of convergence in order to provide a general overview of this phenomenon and highlight functional similarities that would be missed if the analysis was restricted to identical active sites.

Our analysis shows that convergent evolution of enzymes is not a rare phenomenon, especially when one considers very large enzymatic families and reactions that form the basis of the biochemical working of the cell. The catalytic machinery that has evolved the greatest number of times is the catalytic triad of serine proteases. Active sites using this reaction strategy have been identified in 23 SCOP superfamilies. Even though such a high occurrence might be considered exceptional acyltransferase, phosphotransferases, nucleotidyltransferases and glycosidases all contain active sites that have evolved four or more times.

Transformational analogues have been identified in ~5% of the EC numbers present in CSA; in about half of the cases the unrelated enzymes also use comparable mechanisms. This analysis has also been extended to the all the enzymes classified in SCOP. Interestingly the proportion of EC numbers that have transformational analogues remained essentially the same. We may therefore expect this ratio to remain fairly stable as databases grow.

4. REFERENCES

1. Porter CT, Bartlett GJ, Thornton JM. 2004. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* **32**, D129-33

2 Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. 2004. Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* **32**,D226-9

3 Ausiello G, Via A, Helmer-Citterich M. (2005). Query3d: a new method for highthroughput analysis of functional residues in protein structures. *BMC Bioinformatics* **6 Suppl 4**, S5

Dasty2, a web client for visualizing protein sequence features

Rafael C. Jimenez [1,2*#], Antony F. Quinn [1#], Alberto Labarga [1],
Kieran O'Neill [2], Alex Garcia [2], Daniel Jacobson [2], Henning Hermjakob [1]
[1] European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, CB10 1SD, UK
[2] NBN Central, 2nd Floor, Bytes Technology Building, Raapenberg Road, Pinelands, 7405, Cape Town
*To whom correspondence should be addressed: rafael@ebi.ac.uk
# These authors contributed equally to this work.

1. INTRODUCTION

We present Dasty2 (1), a web client for visualizing protein sequence feature information using the Distributed Annotation System (DAS) (2,3). The client establishes connections to a DAS reference server to retrieve sequence information and to one or more DAS annotation servers to retrieve feature annotations (4). It merges the collected data from all these servers and provides the user with a unified, aesthetically pleasing, effective view of the sequence-annotated features. Dasty2 uses AJAX to deliver highly interactive graphical functionality in a web browser by executing multiple asynchronous DAS requests.

Dasty2 has two main modules, retrieval and visualization, as shown in Figure 1. The retrieval module works on the server side. It retrieves annotations from DAS servers and makes them available to the visualization module. The visualization module resides on the client side and facilitates the visualization of the information obtained from the retrieval module.

The retrieval module is a CGI program written in Perl. It can be ported to other programming languages to ease integration into other web applications. The retrieval module is used by Dasty2 but can also be used independently to retrieve XML files from different sources. The CGI program can be executed in any web browser by specifying query parameters in the URL field. It can retrieve feature and stylesheet files from annotation servers, sequence files from reference servers as well as a list of available DAS servers from the DAS registry (5).

The visualization module is based on AJAX. It incorporates standards-based presentation using HTML and Cascading Style Sheets (CSS), dynamic display and interaction using the Document Object Model (DOM), asynchronous data retrieval using the XMLHttpRequest class and JavaScript binding everything together.

The visualization module allows asynchronous loading of information so that protein feature annotations can be displayed as soon as annotation from the first DAS server is provided. Once the information has been retrieved by the client, it is cached, eliminating the need for further server requests for most visual manipulations, such as sorting by feature. The asynchronous loading and local caching of results enabled by using AJAX improve usability and system response time (6).

Technically, Dasty2 is web-browser compatible, lightweight, independent from third party software, easy to integrate into other web based systems, efficient when loading and manipulating annotations, highly configurable and customizable, and interactive and intuitive for users. By being web-based, it is more readily accessible to biological researchers, as they do not need to install specialized software to run it (7). Being JavaScript-based, it is easy for developers to integrate it into their own web systems. Customizability aids this. The use of AJAX improves efficiency when displaying information. Finally, because of its independence from large, complicated third-party libraries and its modularity, Dasty2 is easy to extend.

Visually, Dasty2 provides numerous advantages over other DAS clients (8). Space is used effectively so there is no need to scroll in portions of the screen. It makes use of the standard colours employed by Uniprot and SRS for annotations. Dasty2 uses colours, borders, complimentary shades and separating lines to contrast features with the background and to make the relationships among annotations clear (9). Finally, Dasty2 allows grouping and sorting of annotations by various properties, and zooming within a protein sequence.

Currently DAS annotations do not follow a standard categorization, this makes interpretation and comparison of protein features difficult (10). DAS is adopting the new Biosapiens Protein Annotation Ontology to standardize protein annotations. The corresponding functionality is being developed in Dasty2 to make extensive use of this ontology allowing the client to be able to display and manipulate biologically related data.

2. FIGURES

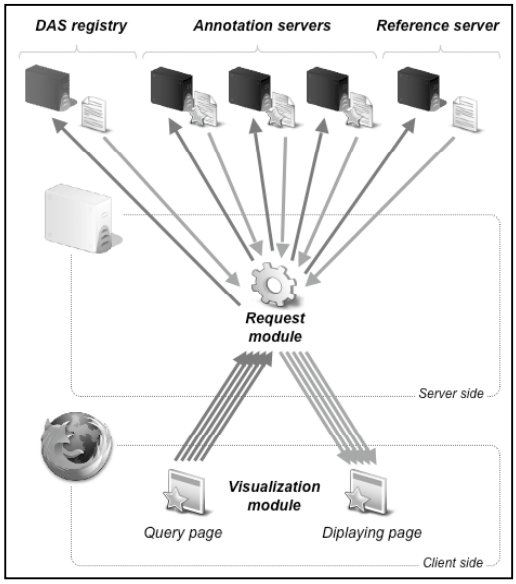


Figure 1: The architecture of Dasty2.

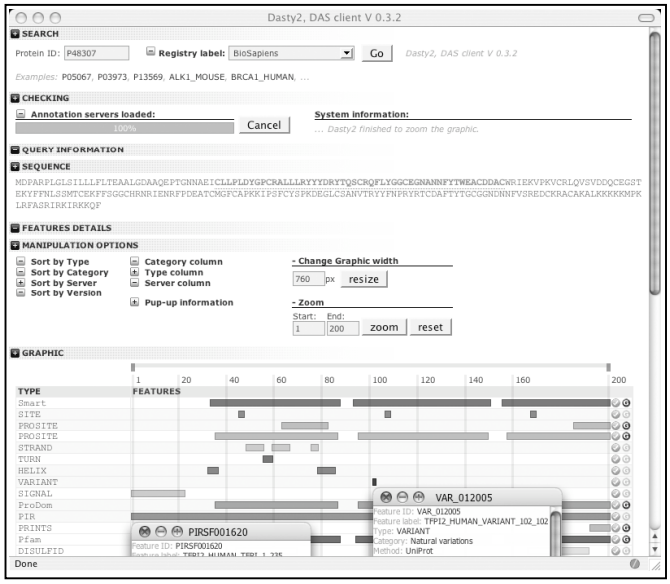


Figure 2 Dasty2, features for Uniprot-ID P4307.

Figure 1. The AJAX based visualization module sends requests to be forwarded on to either the DAS reference server, for discovering DAS sources, or to annotation servers, for retrieving annotations. The request module mediates between the AJAX and the DAS sources.

Figure 2. The configuration file of Dasty2 allows the administrator to configure the layout and visibility of the different sections displayed in the client. Dasty2 enables users to expand and collapse sections, zoom and resize the graphic, display extra feature or server information, drag and drop feature lines up and down throughout the graphic, highlight features and sort features by various properties. Most of these capabilities can be disabled by an administrator if a more lightweight form of Dasty2 is required.

3. REFERENCES

1. Dasty, http://www.ebi.ac.uk/dasty/

2. Dowell, R., Jokerst, R.M., Day, A., Eddy, S.R. & Stein, L. 2001. The Distributed Annotation. *System BMC Bioinformatics* 2:7

3. Biodas, http://www.biodas.org

4. Jones, P., Vinod, N., Down, T., Hackmann, A., Kahari, A., Kretschmann, E., Quinn, A., Wieser, D., Hermjakob, H. & Apweiler, R. 2005. Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics, Oxford Univ Press* 21, 3198-3199

5. DAS registry, http://www.dasregistry.org

6. Paulson, L. 2005. Building rich web applications with Ajax. *Computer* 38, 14-17

7. O'Reilly, T. 2006. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *O'Reilly Media, Inc.* http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

8. Card, S.; Mackinlay, J. & Schneiderman, B. 1999. Readings in Information Visualization: Using Vision to Think. *Morgan Kaufmann*

9. Mullet, K. & Sano, D. 1996. Designing Visual Interfaces. *ACM SIGCHI Bulletin, ACM Press New York* 28, 82-83

10. Reeves, G.A., Montecchi-Palazzi, L., Hermjakob, H. & Thornton, J.M. 2007. The BioSapiens Protein Annotation Ontology. *Communication, Joint AFP-Biosapiens SIG meeting.*

Geometry-inspired Optimization Methods for Structural Motifs for Protein Function Prediction

Brian Y. Chen[1], Drew H. Bryant[2], Viacheslav Y. Fofanov[3], David M. Kristensen[4], Mark Moll[1], Marek Kimmel[3], Olivier Lichtarge[4], Lydia E. Kavraki[1]*

[1] Department of Computer Science, Rice University, Houston, TX 77005, USA
[2] Department of Bioengineering, Rice University. Houston, TX 77005, USA
[3] Department of Statistics, Rice University, Houston, TX 77005, USA
[4] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston TX 77030, USA

*To whom correspondence should be addressed: kavraki@cs.rice.edu

1. INTRODUCTION

One strategy for function prediction is to search the structures of "target" proteins with unknown function for sites which are geometrically and chemically similar to "motifs" representing a known active site. Like all function prediction strategies, the above strategy may have some inaccuracies, such as in the design of the motifs, which may have geometric and chemical dissimilarities to functionally related proteins (not sensitive), or similarities to functionally unrelated proteins (not specific). In this abstract we describe two techniques to optimizing structural motifs so as to increase specificity while maintaining sensitivity. Both techniques are based on a general principle called "Motif Profiling". It is assumed that a reasonably designed motif will be optimized. The presented methods can be used as a post-processing step with many motif design methods.

2. METHODOLOGY AND RESULTS

Improving motif specificity requires the elimination of false positive matches which occur by random chance between a given motif and any large set of target structures. We have developed a general technique called Motif Profiling which provides a measure which seems to be useful for improving specificity in several applications. A "Motif Profile" is a frequency distribution which indicates the frequency of matches observed between a given motif at a specific degree of similarity. In our work the Least Root Mean Squared Distance (LRMSD) is used as a measure of geometric similarity for two sets of atoms that are chemically similar, and motif profiles are explicitly computed by matching a given motif to a representative subset of the PDB (1,2). Due to chance, functionally unrelated sites can still have a low LRMSD to a motif. During Motif Profiling, we aim to skew the distribution of matches such that this is less likely to happen. The threshold distance that best separates functionally related matched from unrelated matches varies per motif. So instead we use the p-value to determine the statistically significant matches (1,2).

The first method we have developed for optimizing, or refining, a given reasonable motif is called Geometric Sieving (GS). The goal of GS is to select a subset from a given motif that has increased specificity while maintaining the sensitivity of the original motif (3). GS takes as input a motif and k, an expected number of motif points in the output motif. It then finds a motif of size k that maximizes the median of the LRMSD of the matches between the motif and a representative subset of the PDB, or in other words a motif with the greatest overall dissimilarity to the PDB. We have shown that this increases the LRMSD of negative matches significantly more than the LRMSD of the positive matches, thus improving the specificity of the motif. To compute the median it is not necessary to compute the full distribution of LRMSD values. Instead, a narrow range for the median can be computed with high confidence with a relatively small number of samples. In (3) we showed that candidate motifs from six well-studied proteins, including a-Chymotrypsin, Dihydrofolate Reductase, and Lysozyme, can be optimized with GS to motifs that are among the most sensitive and specific motifs possible for the candidate motifs.

We also applied Motif Profiling towards the refinement of cavity-aware motifs. The later motifs employ C-spheres to eliminate false positive matching with targets that have atoms occupying volumes essential for protein function. C-spheres are spheres that are rigidly associated with some of the motif points. For a valid match, the C-spheres do not intersect any protein atoms. One difficulty in the design of cavity-aware motifs, in addition to the selection of points for the motif, is the desire to select C-spheres which eliminate many false positive matches. Our method, Cavity Scaling (CS), measures the change in motif profiles as C-spheres expand (4). We observed that some C-spheres, called high-impact C-spheres hereafter, eliminate many false positive matches, and cause the motif profile to shift dramatically towards higher LRMSDs, while low-impact C-spheres do not. In (4) we demonstrated that CS can be applied to identify high-impact C-spheres leading us to believe that in the absence of expert knowledge, CS can guide the design of cavity-aware motifs to eliminate many false positive matches.

3. DISCUSSION

Multiple studies have established the difficulty of designing sensitive and specific motifs for protein function prediction. Our work suggests that because of the speed of current matching algorithms and availability of computational power, it is possible to further refine exist motifs produced by recent motif design efforts by examining the motif profiles when matched to a representative subset of the PDB. Modifying the design of a motif by changing the way an active site is represented, affects the shape and position of the motif profile. In particular, changes to the motif design which cause the median of the profile to shift towards more dissimilar ranges identify changes which reduce the similarity of the motif to the space of known protein structures. In the two separate instances of GS and CS, we have been able to show that Motif Profiling, which is essentially a geometry-inspired approach, is effective in refining given motifs. Our work suggests that geometric criteria may have a role to play in the design of sensitive and specific motifs for protein function prediction. Our work also suggests that the use of large-scale distributed computation on clusters of processors makes possible the design of novel techniques for motif optimization that can shed insight to the very difficult problem of motif design and complement current efforts on protein function prediction.

5. ACKNOWLEDGMENT

This work has been supported in part NSF DBI-0318415 and DBI 0547695 under a subcontract to Rice University. Additional support is gratefully acknowledged from training fellowships of the W.M. Keck Center (NLM Grant No. 5T15LM07093) to B.C. and D.K. and from a VIGRE Training in Bioinformatics Grant (NSF DMS 0240058) to V.F. Experiments were run on equipment funded by NSF CNS-0454333, CNS-042119 in partnership with Rice University, AMD and Cray.

4. REFERENCES

1. B.Y. Chen, V.Y. Fofanov, D.M. Kristensen, M. Kimmel, O. Lichtarge, and L.E. Kavraki (2005). Algorithms for Structural Comparison and Statistical Analysis of 3D Protein Motifs, *Pacific Symposium on Biocomputing 2005*, World Scientific, Hawaii, January, 334-345.

2. D. M. Kristensen, B.Y. Chen, V.Y. Fofanov, R.M. Ward, A.M. Lisewski, M. Kimmel, L.E. Kavraki, and O. Lichtarge. (2006). Recurrent Use of Evolutionary Importance for Functional Annotation of Proteins Based on Local Structural Similarity. *Protein Science* special section on Automated Function Prediction. 15:1530–1536.

3. B.Y. Chen, V.Y. Fofanov, D.H. Bryant, B.D. Dodson, D.M. Kristensen, A.M. Lisewski, M. Kimmel, O. Lichtarge, and L.E. Kavraki (2006). Geometric Sieving: Automated Distributed Optimization of 3D Motifs for Protein Function Prediction, *Research in Computational Biology: 10th Annual International Conference (RECOMB)*, Venice, Italy, April 2006. Published in Lecture Notes in Bioinformatics, A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds), Springer, LNBI 3909/2006, 500-515.

4. B.Y. Chen, D.H. Bryant, V.Y. Fofanov, D.M. Kristensen, A.E. Cruess, M. Kimmel, O. Lichtarge, and L.E. Kavraki (2006). Cavity-Aware Motifs Reduce False Positives in Protein Function Prediction, *Computational Systems Bioinformatics (CSB)*, Stanford, CA, Series on Advances in Bioinformatics and Computational Biology, Imperial College Press, 4: 311-323.

Diverse data combination for gene function prediction by evolutionary ensembles

Yiannis A.I. Kourmpetis*[1], Ate van der Burgt[2], Marco C. Bink[1], Roeland C.H.J van Ham[2] and Cajo J. ter Braak[1]
[1] Biometris, Wagenignen University and Research Center, 6700 AC Wageningen, The Netherlands
[2] Applied Bioinformatics, Plant Research International, 6708 PB Wageningen, The Netherlandds
*To whom correspondence should be addressed: yiannis.kourmpetis@wur.nl

1. INTRODUCTION

Determination of the function of genes is a major field of research in life sciences. The continuous accumulation of diverse genomic data makes the traditional experimental procedures impractical, bringing out the necessity of computational support. The proper combination of heterogeneous information is a key point for the development of a reliable computational method for gene function prediction. In this study we develop such a method for the efficient integration of multiple and diverse data sources.

2. METHOD

Previous results (1) showed that this problem must be addressed in a multi-label classification context, where the goal is to classify a gene into one or more Gene Ontology nodes that correctly describe its function. An ensemble of Support Vector Machine classifiers is constructed in order to improve the prediction accuracy, but also to integrate efficiently the data sources. Each classifier is trained by a different subset of features selected from all the data sources. These subsets of features are selected under an evolutionary combinatorial optimization framework. Given an instance, the final prediction is made by fusing the outputs of the base classifiers. An overview of the procedure is shown at Fig 1.

3. RESULTS

The performance of our method is evaluated using yeast data (sequence attributes, protein-protein interactions, microarray data and localization data). We also compare it with other strategies, namely early and late integration (2).

4. FIGURES

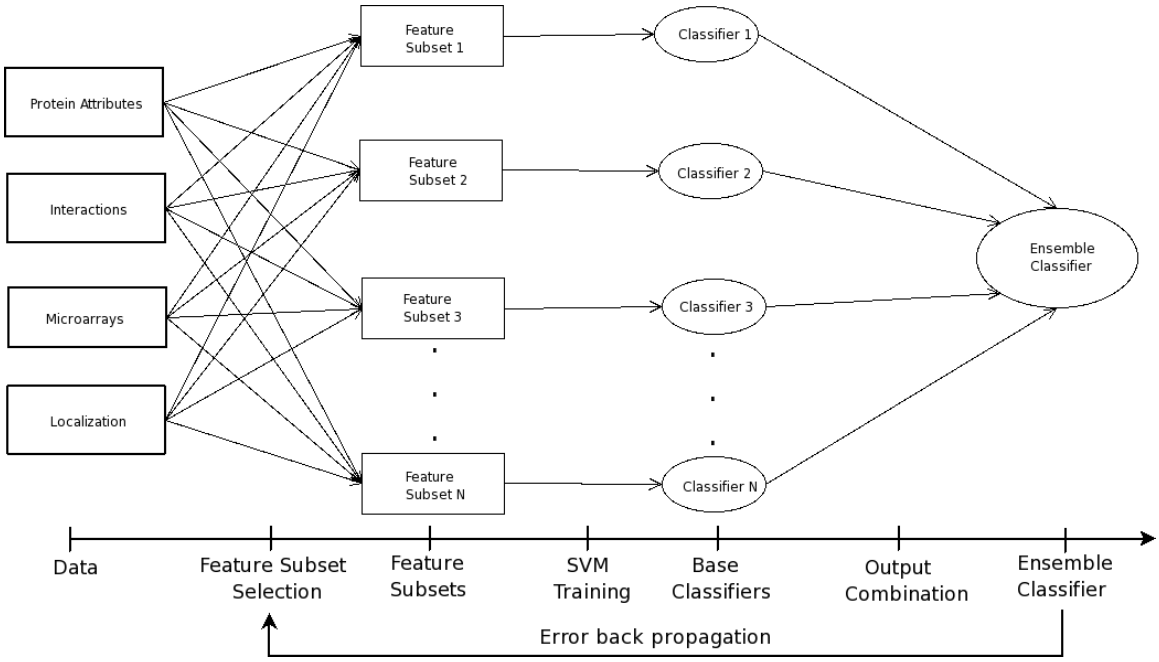


Figure 1. Diverse data integration for gene function prediction by evolutionary SVM ensembles.

5. REFERENCES

1. Kourmpetis *et al* 2007. The use of multiple hierarchically independent Gene Ontology terms in genome annotation and gene function prediction. Submitted.

2. Pavlidis P *et al* 2002. Learning gene functional classifications from multiple data types. J. Comp. Biol. (9) 2. 401-411.

Visual Genomics and Distributed Context Correlation Analysis: Gigantic Palindrome Disintegration as a Common Event of Genomes Evolution, Possible Origin of Function Diversity and Co-functionality

Sergei A. Larionov[1*], Alexander Loskutov[1], Eugeny V. Ryadchenko[1], Maria S. Poptsova[1,2], Ilya A. Zakharov[3]

[1]Physics Faculty, Moscow State University, Moscow, 119992, Russia, [2]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA, [3]Laboratory of Comparative Genetics of Animals, Vavilov Institute of General Genetics, Moscow, 119333, Russia
*To whom correspondence should be addressed: serglarionov@yandex.ru

"..From satellites images to faces recognitions"

1. INTRODUCTION

We discovered that chromosomes of many species, from bacteria to human, have in their origin a gigantic palindrome (palindromes) that disintegrated during evolution process by inversion, duplications, mutation drift and other kind of rearrangements [1,2] . This type of palindromes have length from mega bases to several tens mega bases (it is possible more long) and as a result of large scale rearrangements and drift have only context correlation nature [4,5].

Huge palindromes, from tens kbs up to hundreds kbs long, were known and analyzed during the last several decades [Ford et al, 1986]. Some models of the large palindrome formation in different aspects of medical problem were also discussed [Tanaka et.al, 2005] . Most of them considered as unusual event and very often associated with cancer problem or illness chromosome rearrangement. We found gigantic imperfect palindromes in different chromosomes of many species as a very offen structural element of sequences formation [1]. It should be noted that usually, owing to the complexity of the large imperfect palindrome identification, the regions of the inverted repeats are taken into consideration. One of the large scale approach has been made in the paper [Warburton et al., 2004]. The authors have analyzed this type of the sequence organization for the human genome and found the large number of inverted repeats with the length up to 1Mb. Most of them, for all human chromosomes, contains about 100 nucleotides on the length distribution. But based only on the inverted repeats, it is very difficult to say about a sequence organization and find the complex inversion clusters of the large imperfect palindrome. Very recently, several paper devoted to the analysis of the human gigantic palindromes have been published. In one of them [Bhowmick et al., 2007] the human ampliconic gene families has been analyzed. In the X human chromosome the authors found the twin gene series located on the distance about 15 Mbs, and supposed that this is a palindrome sequence. Another paper [Bansal et.al., 2007] presented ``large inversion polymorphisms" in the human genomes which they found by SNP's from HapMap data project. This type of the data presentation has the same limitations as in [11], but the authors found a large number (far from all) of the gigantic inversion regions up to the 10Mbs length. It is also difficult to say about the evolution and the complexity of these inversions if we rely upon these data. At the same time, it is very important to verify the large scale data. Analysis of the problem of the verification of the large imperfect inversion sites is presented in [Turner et al., 2006], where the authors used PCR technologies on single molecules of the human genomic DNA and experimentally found the inversion breakpoint by repeat-specific markers.

2. METHODS AND RESULTS

Our data obtained thank to a highly visual character of 2D DNA walk method [3] , that allow us to see full chromosome with several hundred mega bases long as unique portrait image [4]. We analyzed gigantic palindromes in detail in different scale and found low sequence similarity by usual sequence alignment methods and large scale context similarity character by additional sequence context separation analysis and found strong correlations between a group of complex sequences sites on opposite positions of imperfect palindromes [2]. We also found a small group of sites with paralog hits within proteins of close and different functions, that located on complementary strands (Watson and Crick) of such palindromes within context correlated sites. Some of them organized in functional clusters, that tested by us in KEGG database. We use 2D walk method as a interface of genomes databases and annotations for detection and analysis functional-structural elements of chromosomes and sequences local properties of such sites of co-functionality . We considered this data in significant examples series and suppose that this results will be useful for wide range of problems: from protein clusters prediction [2,5], and metabolical network organization [1] to a evolutionary modeling [4].

3. FIGURES

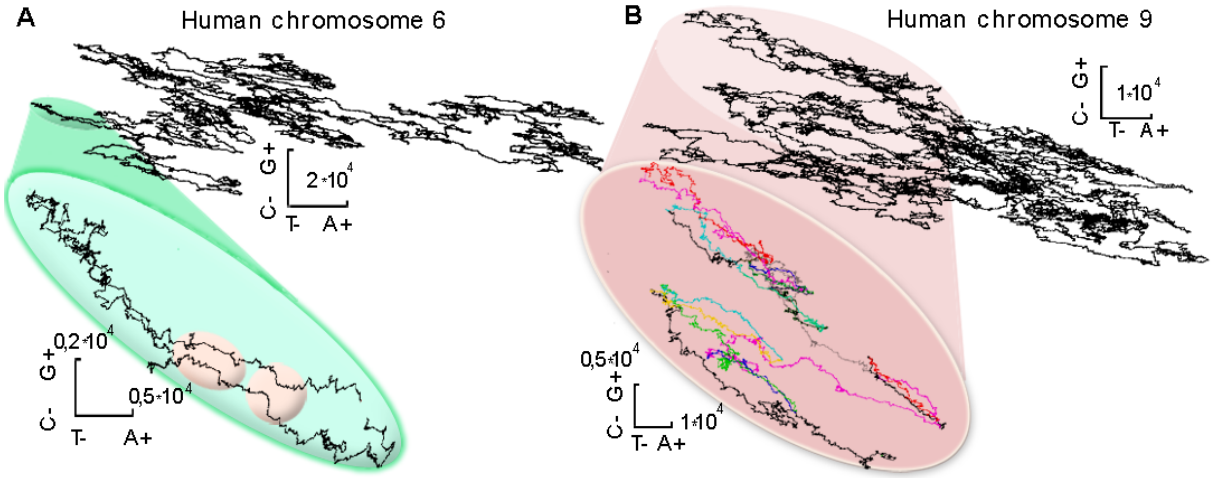


Fig. A. 6 human chromosome (171 mbs): A 3 mbs palindrome (green ellipse) and its one of the most correlated parts (rose ellipses). **Fig.B**. 9 human chromosome (140 mbs): A 30 mbs inverse palindrome region (rose ellipse). Different parts of its inversion are coloured by various colours. 16 mbs of this chromosome are still unknown.

4. REFERENCES

1. Larionov, S.A., Loskutov, A. , Ryadchenko, E.V. 2007. (In Preparation)

2. Larionov, S.A., Loskutov, A. , Ryadchenko, E.V., Poptsova, M.S., Zakharov, I.A. 2007. (In Preparation)

3. Larionov, S.A., Loskutov, A., Ryadchenko, E.V. Genome as a two-dimensional walk. 2005. Doklady--Physics,14, 634--638.

4. Larionov, S.A., Loskutov, A., Ryadchenko, E.V. & Rybalko, S. 2006. Gigantic palindrome diffusion and certain features of genomes evolution. *Proc. of the Int. Symposium on Evolution of Biomolecular Structure*, University of Vienna, Austria, 25-27 May . (Oral presentation and poster session)

5. Larionov, S.A., Loskutov, A., Ryadchenko, E.V. & Rybalko, S. 2006. Visual genomics methods: gigantic palindromes and protein clusters prediction. *Proc. of the Int Symp.: In-silico Analysis of Proteins: Celebrating the 20th Anniversary of Swiss-Prot*, Fortaleza, Brazil, July 30 - Aug. 04 (Short oral presentation and poster session, http://www.swissprot20.org)

6. Ford, M. and Fried, M. 1986. Large inverted duplications are associated with gene amplification. *Cell*, 45, 425--430.

7. Tanaka, H., Bergstrom, D. A., Yao, M. C. and Tapscott, S. J. 2005. Widespread and nonrandom distribution of DNA palindromes in cancer cells provides a structural platform for subsequent gene amplification. *Nature Genet.* 37, 320--327.

8. Bhowmick, B. K. , Satta, Y. and Takahata, N. 2007. The origin and evolution of human ampliconic gene families and ampliconic structure, *Genome Res*. 17,

9. Turner DJ, Shendure J, Porreca G, Church G, Green P, Tyler-Smith C, Hurles ME. 2006. Assaying chromosomal inversions by single-molecule haplotyping. *Nat Methods*. 3(6):427-8.

10. Bansal, V., Bashir, A. and Bafna, V. 2007. Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Res*. 17, 219-230.

11. Warburton, P. E. , Giordano, J., Cheung, F., Gelfand, Y. and Benson, G. 2004. Inverted Repeat Structure of the Human Genome: The X-Chromosome Contains a Preponderance of Large, Highly Homologous Inverted Repeats That Contain Testes Genes . *Genome Res.* 14, 1861-1869.

We would like thank S.Rybalko for his assistance in 2D DNA walk modeling. We also thanks A.Khokhlov, B.Dujon, H.Renauld , P.Schuster, A.Bairoch, L.Hurst, A.Valencia, J.Skolnik and I.Friedberg for useful comments and interest.

firestar and FireDB – a suite for the prediction of functionally important residues

Gonzalo López, Alfonso Valencia and Michael L. Tress*
Structural Biology and Biocomputing Programme, National Cancer Research Centre, Madrid, Spain
*To whom correspondence should be addressed: mtress@cnio.es

1. INTRODUCTION

Genome sequencing projects have lead to a surge in protein sequences lacking experimental functional data and the rise of structural genomics initiatives means that there are many more unannotated structures in the Protein Data Bank (PDB, [1]). Experimental approaches for function characterisation are expensive and difficult to automate and increasingly researchers have turned to computational methods to close the gap between the number unannotated sequences and the number of sequences with known function.

The classical way to overcome this deficit is homology-based transfer of functional annotation, but the transfer of function based solely on the similarity of two sequences is not 100% reliable. There is a need for more sophisticated functional assignment techniques.

Often the most interesting functional information, catalytic residues and ligand binding residues, is to be found at the residue level. Transference of function at the residue level is more laborious since it is necessary to first evaluate the alignments before the interesting residues can be mapped onto the query sequence.

Here we present FireDB [2], the largest inventory of structurally verified functionally important residues, and *firestar* [3], a server for predicting functionally important residues. *firestar* extrapolates from the large inventory of functionally important residues in FireDB and adds information about the local conservation of predicted ligand binding residues. The combined FireDB/*firestar* system (http://firedb.bioinfo.cnio.es) allows the merger of the time consuming tasks of alignment and residue mapping.

These simple procedures include visualisation tools, reliability measures and a choice of pairwise, multiple or structural alignments. Confidence measures for the predictions come from the local sequence alignment reliability measures provided by SQUARE [4].

FireDB

The FireDB database is a databank containing a comprehensive and detailed repository of known functionally important residues. It integrates biologically relevant data filtered from the close atomic contacts in the PDB, crystal structures and reliably annotated catalytic residues from the Catalytic Site Atlas [5]. Ligands that are classified as solvent by mmCif [6] are not included.

PDB sequences are clustered with cd-hit [7] at 97% sequence identity. All functional information is associated to the consensus sequence built for each cluster. FireDB is continuously updated with the growth of the PDB. As of May 6, 2007 FireDB contained a total of 18232 clusters, 10882 of which had associated functional information.

firestar

The server predicts functionally important residues based on the known functionally important residues in FireDB. Queries can be made by protein sequence or structure. Alignments are generated from PSI-BLAST [8] searches of FireDB and functional residues are mapped onto the target sequence via the alignments.

The user has the option of generating further alignments with the multiple alignment program MUSCLE [9] or the structural alignment program LGA [10]. In each case SQUARE evaluates the reliability of the aligned functionally important residues. The binding site residues can be viewed with molecular visualisation tools.

The results are presented in a series of easy to read displays that allow users to compare binding residue conservation across homologous proteins so *firestar* can also be used to evaluate the biological relevance of small molecule ligands present in PDB structures. Conserved sites in two or more homologues implies an evolutionary pressure in residue conservation and suggests biological relevance and *firestar* makes it easier to find homologues with conserved binding sites.

SQUARE

firestar uses a version of SQUARE, developed to predict regions of reliably aligned residues in sequence alignments. The residue scores from SQUARE represent the probability of that a given target residue is aligned according to the evolutionary equivalent template residue. It has been shown that evolutionary conserved binding site residues are almost always involved in ligand binding in the target protein too [11].

References

1. Westbrook J., Feng Z., Chen L., Yang H. and Berman H.M. 2003. The Protein Data Bank and structural genomics. *Nucleic Acids Res* 31:489-491
2. Lopez G., Valencia A. and Tress M.L. 2007. FireDB--a database of functionally important residues from proteins of known structure. *Nucleic Acids Res* 35:D219-223
3. Lopez G., Valencia A. and Tress M.L. 2007. *firestar* - Prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res* in press.
4. Tress M.L., Graña O. and Valencia A. 2004. SQUARE-determining reliable regions in sequence alignments. *Bioinformatics* 20:974-975
5. Porter C.T., Bartlett G.J. and Thornton J.M. 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32:D129-133
6. Bourne P.E., Berman H.M., McMahon B., Watenpaugh K.D., Westbrook J. and Fitzgerald P.M.D. 1997. The Macromolecular Crystallographic Information File (mmCIF). *Meth Enzymol* 277:571-590
7. Li W. and Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659
8. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res* 25:3389-402
9. Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792-1797
10. Zemla A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acid Res* 31:3370-3374
11. Tress M.L., Jones D.T. and Valencia A. 2003. Predicting reliable regions in protein alignments from sequence profiles, *J Mol Biol* 330:705-718

Automated classification of protein structures: a hybrid machine learning approach

Iain Melvin [a], Jason Weston [a], Christina Leslie [b], William Stafford Noble [c*]
[a]NEC Labs of America, Princeton, NJ,
[b]Center for Computational Learning Systems, Columbia University, New York, NY,
[c]Department of Genome Sciences and Department of Computer Science and Engineering, University of Washington, Seattle, WA
[*]To whom correspondence should be addressed: wnoble@u.washington.edu

1 Introduction

Recent protein structural initiatives have rapidly increased the rate of production of new 3D protein structures that need to be annotated and placed in various community database resources. To facilitate the annotation process, we are interested in developing computational methods to automatically assign newly resolved protein structures to structural and functional categories.

Traditional computational methods for comparing protein structures depend on pairwise structural alignment programs such as MAMMOTH [1]. Using pairwise structural comparisons of a query structure against a curated database, one can implement a nearest neighbor strategy to classify the query. More recently, several groups have introduced support vector machine (SVM) methods for the protein structure classification problem. Here, an SVM is trained on both members of a structural/functional class and negative examples using a kernel function for protein structures and learns to discriminate between the two classes; the kernel function may itself use structural alignment scores to represent structures in the classifier. Benchmark experiments have shown that SVM-based discrimination with a MAMMOTH kernel outperforms using MAMMOTH in a nearest neighbor fashion [2]. However, all SVM-based methods are hampered by limited coverage: since a sufficient number of positive examples are need to train the classifier, very small classes that are represented by only one or several structures in the database cannot be covered.

In this study, we develop a hybrid machine learning approach for classifying protein structures into SCOP superfamilies, which are structurally defined classes that are useful for inferring protein function. Our goal is to combine the nearest neighbor method based on structural alignments, which in principle has complete coverage over SCOP, with a higher accuracy but reduced coverage multi-class SVM approach to produce a single full coverage method with overall improved accuracy. The hybrid approach is based on the simple idea of "punting" from one method to another based on a threshold for the predicted class of the current method. We first use a threshold for each class and for each method, all determined by taking a large set of negative training examples (that have not been used to train the classifier in question) and finding a threshold for the classifier that results in a fixed small false positive rate. After punting from the primary method to the secondary method, we also consider different coverage thresholds at which to punt out of the secondary method (i.e. abstain from making a prediction altogether), and we compute error rates of the hybrid method at these different coverage levels. Using this simple punting approach, we find that the hybrid method can consistently outperform the individual component methods at all levels of coverage.

2 Experimental Results

We use structural alignments based on the MASTODON program, a new pre-release version of MAMMOTH, both for the nearest neighbor method and to define a kernel representation for training SVMs to recognize SCOP classes, using an empirical kernel map approach [2]. For simplicity, we used a standard one-vs-all approach for making multi-class predictions from binary SVM classifiers.

We divided the dataset (all of SCOP) into 4 parts: A_{trn}, A_{tst}, B_{trn}, B_{tst}. We determined A_{trn} and A_{tst} to suit the requirement of training and testing binary SVM superfamily classifiers: A_{tst} consists of totally held-out families from superfamilies that have 2 or more member families of at least 3 proteins; A_{trn} consists of all other families belonging to these superfamilies. Dataset B consists of all superfamilies in SCOP that are not covered by dataset A. B is then split into train and test by families at random such that the ratio

Test Set	A_{tst}	B_{tst}	$A_{tst} + B_{tst}$
Test set examples	947	1590	2537
SVM 1-vs-rest [A_{trn}]	17.2	100	51.5
MASTODON [$A_{trn} + B_{trn}$]	34.9	47.8	40.2
SVM→MAST [$A_{trn} + B_{trn}$]	21.2	50.2	33.2
MAST→SVM [$A_{trn} + B_{trn}$]	22.9	48.8	33.6
MAST→C→SVM [$A_{trn} + B_{trn}$]	37.2	47.8	41.6

Table 1: **Superfamily detection error rates for full coverage.** A_{tst} consists of held-out families from superfamilies within the coverage of the SVM classifiers; B_{tst} consists of familes outside of SVM coverage. MAST→C→SVM is a variant of punting where if MASTODON predicts a class that is in the SVM coverage, then we punt to the SVMs; otherwise we make a MASTODON prediction.

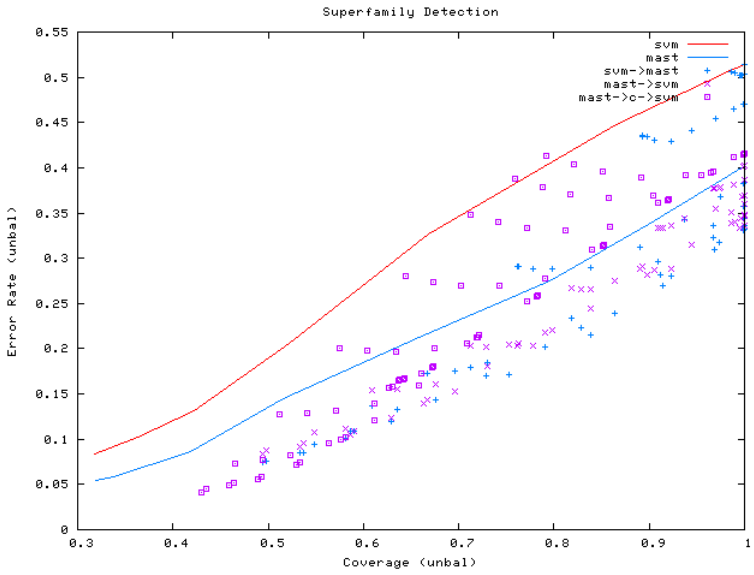


Figure 1: **Unbalanced coverage vs unbalanced error rates for superfamily detection.** The target percentage for both the primary and final punting thresholds are varied in hybrid methods giving a range of coverage and error rates. Starting with either the SVM classifiers as our primary method or with MASTODON, we achieve higher accuracy with the hybrid method. MAST→C→SVM is as in Table 1.

of families for B_{tst}/B_{train} is equal to the ratio A_{tst}/A_{train}. The dataset for superfamily detection has 71 superfamilies in A and 1461 superfamilies in B (total 1532).

We considered punting both from SVMs to MASTODON and vice versa. When using SVMs as the primary method, we used B_{trn} as additional negative examples on which to calculate punting thresholds. We also tried using MASTODON as the primary method and creating thresholds to determine when to punt to the SVM method. In this case, since the nearest-neighbor method had accrued no bias in "training", we used all of the negative superfamilies in A_{trn} and B_{trn} to determine thresholds for each of the superfamilies.

We show results for A_{tst}, B_{tst} and $A_{tst} + B_{tst}$ in Table 1. One can see that the hybrid methods do help in the combined A and B test results, though the error rate at full coverage is still high. Therefore, we included a second level of punting, based on a second set of thresholds to "punt" completely and not give a prediction for an example. Results are shown in Figure 1 and demonstrate improved accuracy at many coverage levels.

References

[1] A. R. Ortiz, C. E. M. Strauss, and O. Olmea. Mammoth (matching molecular models obtained from theory): An automated method for model comparison. *Protein Science*, 11:2606–2621, 2002.

[2] Jian Qiu, Martial Hue, Asa Ben-Hur, Jean-Philippe Vert, and William Stafford Noble. A structural alignment kernel for protein structures. *Bioinformatics*, 2007. In press.

Predicting Gene Ontology terms using an ensemble of calibrated SVMs

Guillaume Obozinski[1], Gert Lanckriet[2], Charles Grant[3],
Michael I. Jordan[1], William Stafford Noble[3*]
[1]Department of Statistics, UC Berkeley, CA, USA [2]Department of Electrical and Computer Engineering, UC San Diego, CA, USA [3]Department of Genome Sciences, University of Washington, Seattle, WA, USA *Address correspondence to noble@gs.washington.edu.

1. Introduction

In this work, we address the problem of predicting Gene Ontology terms from a heterogeneous collection of genomic and proteomic data sets. We treat each term independently, and we train a large collection of binary classifiers that answer the question, "Should this gene be assigned the Gene Ontology term X?" The input to the classifier includes ten different data sets, which were assembled as part of a recent comparative assessment of gene function prediction methods [2], including protein similarity scores across multiple proteomes, protein-protein interaction data, microarray expression data, etc.

Our method uses a collection of support vector machine (SVM) classifiers, trained separately for each term and each data type. The output of each trained SVM with respect to a new gene is a discriminant value, which is proportional to the distance of that gene to the separating hyperplane. This number is not comparable from one SVM to another, and moreover, a discriminant value cannot be interpreted directly as an absolute confidence measure. Therefore, one needs to *calibrate* this value, i.e., map it to a probability that reflects the degree of confidence that we have when assigning that gene to the GO term.

This work focuses on calibration methods. We investigate three different strategies: (1) logistic regression learned from all SVM outputs, (2) a naive Bayes classifier based on fitting Gaussian distributions to the output of each SVM for the positive and for the negative examples separately [3], and (3) a similar method that uses an asymmetric Laplace distributions in place of the Gaussian [1]. The results show that the logistic regression performs best.

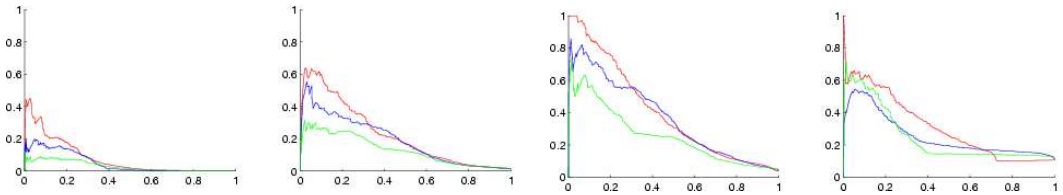
2. Methods

Our method consists of three phases: (1) compute a collection of *kernel matrices* (defined below), (2) train an SVM for each term and each kernel matrix, and (3) combine and calibrate the SVM discriminant scores.

Our first challenge is to choose an appropriate representation of each data type. The SVM relies upon a particular form of similarity function (the *kernel function*) to convert a given data set into an all-by-all similarity matrix (the *kernel matrix*). Thus, choosing a representation is equivalent to choosing a kernel function. Our strategy is to construct several kinds of kernels for each data matrix. These include linear kernels ($K(x,y)=x\cdot y$) with and without normalization, radial basis function kernels ($K(x,y)=e^{-\|x-y\|^2}$), and—for the protein-protein interaction data—diffusion kernels computed on the binary protein adjacency matrix.

The resulting kernel matrices can be combined algebraically, which allows a single SVM to learn from diverse types of data (e.g., networks, sequences and vectors). Techniques exist for learning the relative weights of different data types as part of the SVM optimization; however, these methods are computationally expensive and do not easily handle missing data. Therefore, in the second phase, we instead train a collection of single-kernel SVM classifiers for each GO term.

As mentioned above, the discriminant values produced by an SVM need to be calibrated in order to allow the comparison of discriminants produced by different SVMs and to allow for easier interpretation of the values themselves. We use held-out data to train a regressor that maps from discriminant scores to probabilities. Rather than training separate regressors for each SVM, we train a single regressor to map from a vector of SVM discriminant scores to a single probability. The number of entries in the input vector is equal to the number of kernel matrices that we computed in the first



$3 < n \leq 10$ $10 < n \leq 30$ $30 < n \leq 100$ $100 < n \leq 200$

Figure 1: **Precision-recall curves for three calibration methods.** Each curve is computed using a single, fixed test set which was not used during the training procedure. The methods are logistic regression (red), Gaussian fitting (green) and asymmetric Laplace fitting (blue). These curves are computed only using terms from the cellular compartment Gene Ontology. Plots for the other components are qualitatively similar.

phase. We investigate three strategies for training this regressor: logistic regression trained using the classical iterative re-weighted least squares method, and two methods that fit distributions—either Gaussian or asymmetric Laplace—to the discriminant scores of the positive and negative held-out data.

When applying the logistic regression method, we are faced with a missing data problem, because for most genes, only a subset of the data types are available. We therefore train a collection of 11 regressors, each corresponding to a commonly occurring pattern of missing data. If a test point pattern exactly matches the pattern of one of these regression models, then we calibrate that test point with that model. For the remaining points, we identify the model that matches most closely the observed pattern of missing data, and use imputation to fill in the missing values.

3. Results and discussion

We use precision-recall curves to measure the performance of our classifiers, and we perform this analysis separately for four different sets of GO terms. The splits correspond to the number n of positive examples available for training each classifier ($3 < n \leq 10$, $10 < n \leq 30$, $30 < n \leq 100$, and $100 < n \leq 200$). The results in Figure 1 show that, for all four sets of GO terms, logistic regression provides superior classification performance.

Our group participated in the recent mouse function prediction assessment [2], using the first of the three methods described above. We have subsequently continued to work on the same benchmark. Thus far, we have restricted our analyses to cross-validation testing within the training set that was distributed to all the groups. In addition to the calibration methods described above, we are currently testing a variety of methods for reconciling our GO term predictions with respect to the GO graph. These methods include techniques based on KL-projection, a computationally improved version of a previously described Bayesian network approach, and a conditional random field model. By the time of the workshop, we plan to have completed these analyses and tested our final model on the test set.

4. References

[1] P. N. Bennett. Using asymmetric distributions to improve classifier probabilities: A comparison of new and standard parametric methods. Technical Report CMU-CS-02-126, Carnegie Mellon University, Pittsburg, PA, April 2002.

[2] L. Peña-Castillo, (37 *authors omitted*), and F. P. Roth. A critical assessment of *M. musculus* gene function prediction using integrated genomic evidence. Submitted, 2007.

[3] J. C. Platt. Probabilities for support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

The DIMA Web Resource on Domain Interactions

– New Data and Features

Philipp Pagel [1,2*], Matthias Oesterheld [2], Volker Stümpflen [2] and Dmitrij Frishman [1,2]
[1] Lehrstuhl für Genomorientierte Bioinformatik, Technische Universität München, Wissenschaftszentrum Weihenstephan, Am Forum 1, 85350 Freising, Germany
[2] Institut für Bioinformatik / MIPS, GSF–Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
*To whom correspondence should be addressed: p.pagel@wzw.tum.de

1. INTRODUCTION

Conserved protein domains carry many of the functional features found in the proteins of an organism. Catalytic activity, binding to other proteins, small compounds or DNA as well as structural properties can often be attributed to distinct domains. While many protein-protein interactions are highly protein specific, others are mediated by interaction domains which in turn bind to short motifs or other domains. In addition, functional relationships among proteins which do not necessarily involve physical contact constitute another form of interactions which often involve conserved domains.

A wealth of experimental data is available on functional and/or physical interactions among proteins. Large protein interaction databases allow convenient access to carefully curated data from individual experiments as well as high-throughput endeavors. On the level of protein domains on the other hand, very little such data is available emphasizing the importance of in silico methods to predict domain-domain relations.

Based on the well known method of phylogenetic profiling we initially implemented a method to predict domain interactions from their pattern of evolutionary co-occurrence in different organisms (1). These predictions used the excellent PFAM domain database and models (2). With the original version of the Domain Interaction MAp (DIMA) web resource we allowed easy access to the method for others (3). Since the original publication, many features and a lot of new data have been added to DIMA turning it into a comprehensive resource on domain-domain interactions.

To complement the domain profiling, two sources of experimental data were included: iPfam (4) is a database of domain-domain contacts observed in solved protein structures in PDB (5). 3did (6) represents a resource following the same idea independently and with a slightly different approach. These experimental data sources represent a great addition to our predictions and serve as yardsticks for validation.

An other important approach to domain-domain interaction prediction tries to identify the most likely domain combinations to mediate observed protein-protein interactions. Many variations and improvements on this idea can be found in the literature (7-9). In DIMA, we have recently included the method by Riley et al. (9) which, in our opinion, represents the most advanced approach of this kind, today. In addition to experimental protein-protein interaction data, we also apply this algorithm to predicted protein interactions from the STRING (10) database. Although this does not yield the good predictive power as predictions based on experimental data, the results are clearly good enough to provide a valuable addition – especially with respect to increased coverage.

DIMA provides the only "one-stop shopping" resource integrating all the above mentioned methods and data

sources. With the upcoming new release, we are also substantially upgrading the data used by each prediction method. E.g. domain profiling will be based on over 400 completely sequenced genomes represented in the PEDANT genome database. At the same time, we are working on the implementation of yet other prediction approaches to further complement the selection of methods.

In summary, the DIMA domain-interaction resource is a steadily growing and actively maintained database which provides an easy to use web interface for interactive use and also allows bulk generation of entire domain-interaction networks based on the users preferences. We hope and believe that it represents a valuable tool for researchers interested in aspects of domain function and the relations among conserved protein domains.

2. REFERENCES

1. Pagel, P., Wong, P. and Frishman, D. 2004. A Domain Interaction Map Based on Phylogenetic Profiling *Journal of Molecular Biology* 344(5): 1331-1346

2. Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C. and Eddy, S. R. 2004. The Pfam protein families database. *Nucleic Acids Res 32*, D138-D141

3. Pagel, P., Oesterheld, M., Stümpflen, V, Frishman, D. 2006. The DIMA web resource – exploring the protein domain network. *Bioinformatics* 22(8): 997-998

4. Finn, R. D., Marshall, M. and Bateman, A. 2005. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 21: 410-412

5. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. 2000. The Protein Data Bank. *Nucleic Acids Research* 28: 235-242

6. Stein, A., Russell, R. B. and Aloy, P. 2005. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Research* 33: D413-D417

7. Deng, M., Mehta, S., Sun, F. and Chen, T. 2002. Inferring domain-domain interactions from protein-protein interactions. *Genome Research* 12, 1540-1548

8. Sprinzak, E. and Margalit, H. 2001. Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology* 311: 681-692

9. Riley, R., Lee, C.; Sabatti, C. and Eisenberg, D. 2005. Inferring protein domain interactions from databases of interacting proteins. *Genome Biology* 6: R89

10. von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. and Bork, P. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* 33: D433-D437

Improving Function Prediction with Sequence Alignment Network

Keunwan Park†, Wonchul Lee†, Dongsup Kim*
Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea 305-701
*To whom correspondence should be addressed: kds@kaist.ac.kr
†The authors wish it to be known that the first two authors should be regarded as joint First Authors.

1. INTRODUCTION

As the number of proteins with unknown function increases, it becomes important to predict functions of those proteins before the actual functions are revealed. The basic and standard method used for this matter is Basic Local Alignment Search Tool (BLAST) (1). It is useful in inferring functional and evolutionary relationship between protein sequences. However, BLAST often fails to identify the function of a query when there is no significant hit. In this study, we develop a noble method which improves the function prediction accuracy by BLASTP search result including sub-alignment information, which leads to the construction of the sequence alignment network. Our method shows better performance than PSI-BLAST (2) even when there is no significant hit (e-value$\geq$10).

We assume that [1] Even proteins of similar function whose sequence similarity is low are subject to share certain aligned regions, [2] If two proteins of similar function are aligned to the similar region of a query, it is likely that the query functions similarly with those two proteins. [3] Sub-alignments can possess weak signal, so they give functional clues.

To make pre-compiled sequence alignment network, we performed one-against-all BLASTP search using 65% sequence homology-reduced Swiss-Prot database (3) with EC (4) annotated (33,584 sequences). Each sequence in the database becomes a node of the network. For the edges and their strength, each sequence is seeded and searched against the database with e-value threshold 10,000 which guarantees sub-alignments appearance in the result. Then, edges are generated between the seed protein and all the reported proteins with bit scores as their strength. Note that there can be more than one edge between any two nodes when those two proteins have sub-alignments. Once the alignment network is completed, the function annotation to an unknown query becomes possible. After scanning through the network with a query, local structures of the network, so called "Triplet" are found. Triplet structure (shown in Figure1 right) consists of a query and other two proteins in the network sharing similar function. These sequences in the triplet should have overlapped alignment regions. On the other hand, we also need to define a function-reliability score of the triplet, because many triplets can exist for a certain alignment region of the query. We denote function-reliability score (S) as :

$$S = bitscore(Q, A) + bitscore(Q, B) + R(A, B)\text{, where}$$

bitscore(,) : bit score between two proteins

$$\text{R(A,B)}: \quad a * \exp\left(\frac{-|bitscore(A, B) - b|}{c}\right).$$

We empirically set the scale parameter "a" to 30, base "b" to 40 (corresponding e-value 0.01), and the deviation "c" to 70. Gaussian function is used when calculating the score between A and B, because A and B lose their importance when they are either extremely close homologs or sequentially-unrelated proteins. Finally, the best scoring triplet is found, and the function of proteins within the triplet is assigned to the query.

Our method is based on the BLASTP result with the loose e-value threshold (10,000) to encompass either sub-alignments or weak alignments in terms of significance level, but possess functional hints. For the performance evaluation, we randomly chose 1000 sequences in the network as test queries. Then, each test query was searched against the network excluding the query itself. We compared our result with PSI-BLAST in two ways. First, we checked if the function of the best hit accords with the query's function while changing the e-value threshold from $1e^{-200}$ to 10 (Figure2-(a)) and from 10 to 100 (Figure2-(b)). The functional match is defined as upper 3-digit agreement of the EC numbers. As seen, our method performs better than PSI-BLAST in both ranges. We also could confirm the effect of "Triplet" by the result of all the

protein-pairs sharing function (Figure2-(b) dotted line). Second, we measured how percentage of proteins have similar function with the function of the query when the top 30 are considered (Figure2-(c)). It clearly shows that our method places the proteins sharing function with the query in the high-rank position well.

In conclusion, our method successfully improves PSI-BLAST with the sequence alignment network, and proves the availability of sequence alignment information including sub-alignments in function prediction. The result shows that our approach performs better than PSI-BLAST both when significant hits exist and when only weak hits are available. On the other hand, our method can be improved if more sub-alignment information is available by dynamic programming and if the better scoring scheme is designed.

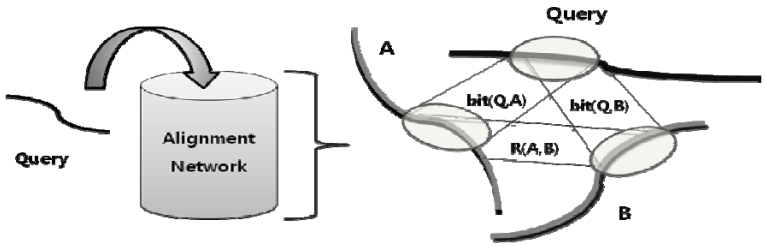


Figure1. Function Annotation Process. Given an unknown query, it is scanned through the alignment network, and triplet structures are found. Each triplet is scored based on the bit scores of the sequences within it. Finally, the function of protein A and B within the best scoring triplet is transferred to the query.

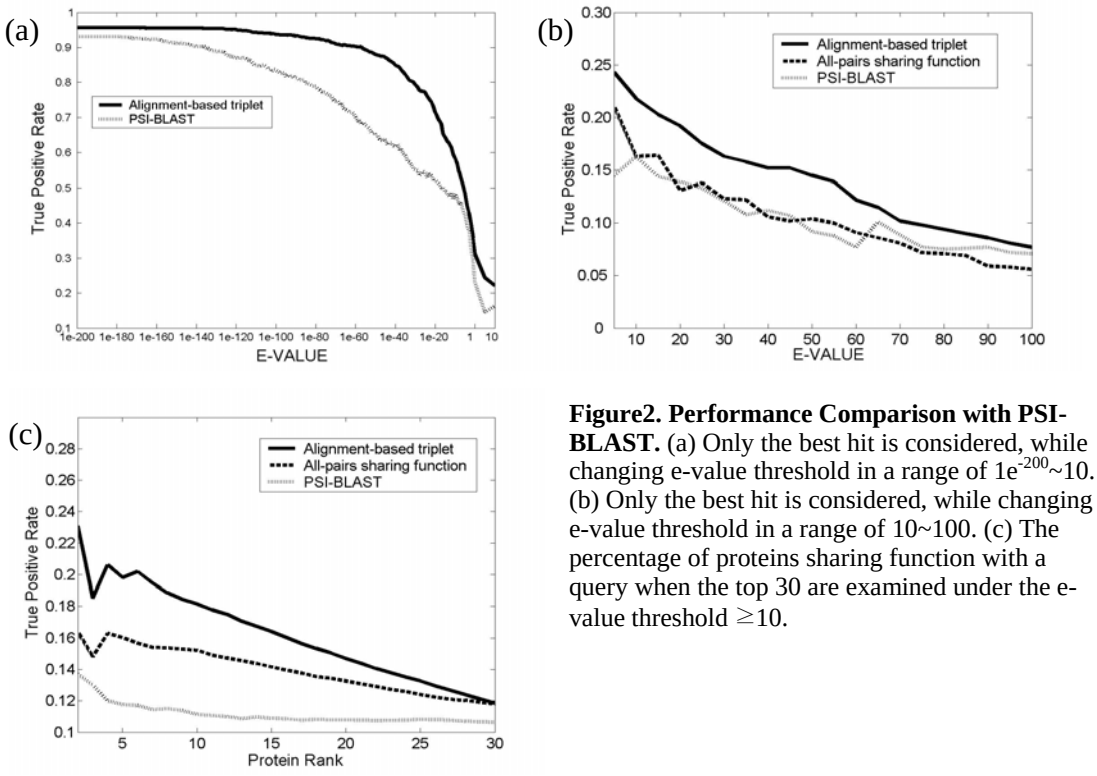


Figure2. Performance Comparison with PSI-BLAST. (a) Only the best hit is considered, while changing e-value threshold in a range of $1e^{-200}$~10. (b) Only the best hit is considered, while changing e-value threshold in a range of 10~100. (c) The percentage of proteins sharing function with a query when the top 30 are examined under the e-value threshold $\geq$10.

2. REFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. *J Mol Biol* . 215(3):403-410.
2. Altschul SF, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman DJ. 1998 Gapped blast and psi-blast: a new generation of protein database search programs. *FASEB Journal.* 12(8), pp. A1325.
3. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003 ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research.* 31(13):3784-3788.
4. Bairoch A. 2000 The ENZYME database in 2000. *Nucleic Acids Research.* 28(1):304-305.

Prediction of subcellular localization in eukaryotes at the basis of large-scale genome annotation

Andrea Pierleoni, Pier Luigi Martelli, Piero Fariselli and Rita Casadio
Biocomputing Group, Dept. of Biology University of Bologna via Irnerio 42, 40126 Bologna, Italy

ABSTRACT
In this work we present an integrated platform for large-scale eukaryotic genome annotation based on the prediction of subcellular localization, GPI-anchor prediction and membrane protein discrimination into inner and outer classes.

Large scale proteomic projects have determined a huge number of aminoacidic sequences whose functions are, in the largest part, still unknown. In eukaryotes compartmentalization plays a major role in intracellular biochemical pathways. However the determination of subcellular localization with experimental high-throughput procedures is a difficult task and computational procedures are needed.
We developed BaCelLo (1), a predictor for five classes of subcellular localizations (secretory pathway, cytoplasm, nucleus, mitochondrion and chloroplast) that is based on different SVMs organized in a decision tree. The system exploits the information derived from the aminoacidic sequence and from the evolutionary information contained in alignment profiles. It analyzes the whole sequence composition and the compositions of both the N- and C-termini. The training set is curated in order to avoid redundancy. For the first time a balancing procedure is introduced in order to mitigate the effect of biased training sets. Three kingdom-specific predictors are implemented: for animals, plants and fungi, respectively. When distributing the proteins from animals and fungi into four classes, accuracy of BaCelLo reach 74% and 76%, respectively; a score of 67% is obtained when proteins from plants are distributed into five classes. BaCelLo outperforms the other presently available methods for the same task and gives more balanced accuracy and coverage values for each class. BaCelLo is also described in Nature Protocols, in the Bioinformatics section (2)
BaCelLo can be accessed at http://www.biocomp.unibo.it/bacello/.
BaCelLo is currently under integration in a workflow which will allow GO functional integration, prediction of GPI-anchors and discrimination between inner and outer membrane proteins. The workflow will be tested on large-scale genome annotation.
With a suite of machine learning based methods, developed in house (BaCelLo, SpepLip (3) and ENSEMBLE (4)), we presently built eSLDB (eukaryotic Subcellular Localization DataBase) (5) an online database collecting the annotations of subcellular localization of eukaryotic proteomes. So far five proteomes have been processed and stored: Homo sapiens, Mus musculus, Caenorhabditis elegans, Saccharomyces cerevisiae and Arabidospis thaliana. For each sequence, the database lists localization obtained adopting three different approaches: 1) experimentally determined (when available); 2) homology based (when possible); 3) predicted. All the data are available at the website and can be searched by sequence, by protein code and/or by protein description. Furthermore a more complex search can be performed combining different search fields and keys. All the data contained in the database can be freely downloaded in flat file format.
The Database is available at: http://gpcr.biocomp.unibo.it/esldb/

1. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) *BaCelLo: a Balanced subCellular Localization predictor.* Bioinformatics, **22**, e408-e416.
2. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) *BaCelLo: a Balanced*

subCellular Localization predictor. Nature Protocols **DOI:** 10.1038/nprot.2007.165

3. Fariselli,P., Finocchiaro,G. and Casadio,R. (2003) *SPEPlip: the detection of signal peptide and lipoprotein cleavage sites.* Bioinformatics, **19**, 2498-2499.
4. Martelli,P.L., Fariselli,P. and Casadio,R. (2003) *An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins.* Bioinformatics, **19**, i205-i211.
5. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) *eSLDB: eukaryotic Subcellular Localization DataBase.*, Nucleic Acids Res., in press.

The BioSapiens Protein Annotation Ontology

Gabrielle A. Reeves*, Luisa Montecchi-Palazzi, Henning Hermjakob and Janet M. Thornton

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.
* To whom the correspondence should be addressed: gabby@ebi.ac.uk

1. INTRODUCTION

Currently, the BioSapiens Network of Excellence has produced annotations at 19 partner sites providing 69 different distributed annotation sources. This comprises information for genomic sequences and protein sequences as well as for protein structures. Visualisation of these data is provided by three DAS clients; the Ensembl genome browser for both genomic and proteomic annotations, Dasty2 for protein sequence annotations and Spice for both protein sequence and structural annotations. These three clients are accessible from the BioSapiens web portal (http://www.biosapiens.info). In the future, the BioSapiens portal will provide a valuable central resource for many annotations. The power of this resource derives from combining and comparing annotations from many sources, which relies on the organisation and presentation of the data displayed. Currently, annotations are free text which makes clustering, interpretation and comparison of biologically 'related' data difficult. In order for biological inferences to be made by the user, it would help if the data tracks were categorised so that 'like' annotations can be more easily compared and duplications in the data identified. Figure 1 shows a screen shot from Dasty2. Two of the servers, UniProt and PDBSum annotate residues in contact with a copper ion and in the present format it is difficult to infer biological meaning from this layout. The controlled vocabulary will also allow identification of duplications in the data, for example, both the UniProt and Interpro servers have provided annotations of the same PROSITE domain.

In order to carry out this categorisation, we have created a protein annotation ontology, thus providing a means of clustering all annotations into a biologically meaningful manner but also providing a controlled vocabulary as a means of standardizing the language used in these annotations. The ontology is available at http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=BS and is divided into "positional" and "non-positional" annotations. The positional annotations are those annotations which can be located to a specific region of sequence, for example, domain or active site (Figure 1). The non-positional annotations are those which are related to the whole sequence such as "publication" or a description of the general function of the protein

2. FIGURES

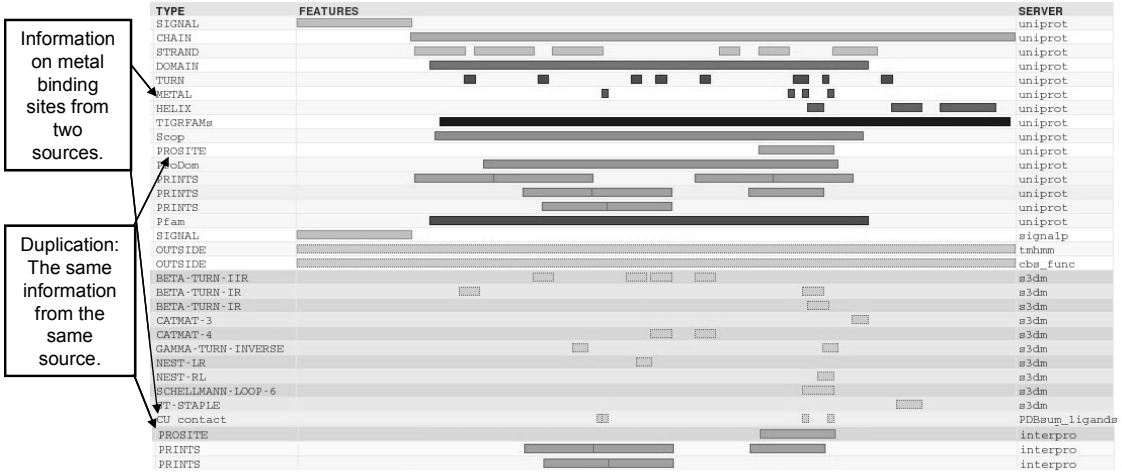


Figure 1. A screenshot of the positional annotations for a domain in the cupredoxin superfamily (CPC_CUCSA) viewed in the Dasty2 client.

Multi-stranded bioinformatics to investigate a neglected superfamily

Daniel J. Rigden*
School of Biological Sciences, University of Liverpool, Liverpool, UK. L69 7ZB
*To whom correspondence should be addressed: daniel@liv.ac.uk

1. INTRODUCTION

The accurate, automatic functional annotation of protein sequences in the genomic era represents a huge challenge to the bioinformatics community [1]. As sequences and structures accumulate the concept of the superfamily is becoming ever more important. Members of superfamilies have a common, but distant, evolutionary ancestor so that their shared homology may be difficult to detect and they may have evolved diverse novel functions. Modern profile and fold recognition methods have addressed the homology detection issue but the often sparse experimental sampling of functions in superfamilies poses twin questions to bioinformaticians – how many members of the superfamily can be reliably annotated with experimentally determined functions, and can predictions of novel functions be made for the remaining members?

The His-phosphatase/mutase superfamily illustrates these issues very well. It is very large – 3300 unique sequences – and, although the first characterised enzyme was cofactor-dependent phosphoglycerate mutase (dPGM), largely contains phosphatases of impressively diverse substrate specificities and wide-ranging physiological roles. Although experimental functions are now being determined at an accelerating pace, the paucity of information available in the 80s and 90s led to widespread misannotation by automatic pipelines, exacerbated by the creation of mislabeled COG entries. The time is ripe for a reappraisal and reannotation of the family.

2. RESULTS

As a first step towards the systematic dissection of the superfamily we have used CLANS [2] to cluster members of a non-redundant database (Figure 1). The results of an initial partition show that well-defined clusters correspond well to known activities: where clusters contain known activities and crystal structures provide information about specificity determinants, all members can be confidently assigned the corresponding function.

Our grouping showed several well-defined groups with no known activities. We are adopting several bioinformatics approaches to tentatively predict functions for these. One of the most informative techniques has been a domain association scan using RPS-BLAST and the CDD database [3]. Unfortunately, manual sifting of these results was required to remove erroneous hits in situations where genome annotation had incorrectly combined sequences from separate but neighboring genes. Nevertheless, a number of reliable, functionally suggestive combinations were found. In group 18, for example, the His phosphatase/mutase domain follows a NUDIX domain, most likely of the MutT activity which aids replicative fidelity by hydrolysing mutagenic 8-oxo-dGTP to 8-oxo-dGMP and PPi. This suggests that the His phosphatase/mutase domain may catalyse the further degradation of 8-oxo-dGMP to 8-oxo-deoxyguanosine.

We are also exploring structure modeling, previously used to successfully predict specificity in this superfamily [4], as well as non-homology approaches and mapping of group taxonomic distributions in order glean clues as function. Group 20 shows how synthesis of results from multiple bioinformatic approaches yields functional insights. Its presence exclusively in gram negative bacteria led subcellular localization predictors to strongly indicate periplasmic location. A structural model of a group representative (Figure 2) showed that, as well as having a full complement of catalytic residues, this group, like SixA and AfrS, lacks the large domain insertion otherwise ubiquitous in the superfamily. This characteristic leads to a much larger, more open and planar substrate binding site than the cavity found in dPGM and most phosphatases. As such, it is strongly associated with very large substrates, those so far characterized being exclusively

phosphorylated receiver domains of 2-component signaling systems. Single members of the group showed two suggestive domain combinations, the first being a PCC domain, otherwise found only in extracellular proteases and the second resembling a CAM-dependent protein kinase alpha subunit. These domains have roles in protein binding and multimer assembly, respectively, with the former strongly suggestive of a protein substrate. Thus, group 20 appears to comprise a further family of protein phosphatases involved in regulation of 2-component signaling pathways.

In summary, this ongoing work is yielding significant functional insight into the neglected His-phosphatase/mutase superfamily. It is becoming ever clearer that physiological roles are far more diverse than previously suspected. Notably, while certain predictions are or could be integrated into automated systems, there is still an important role for experts with specialist knowledge of particular proteins, families or superfamilies.

3. FIGURES

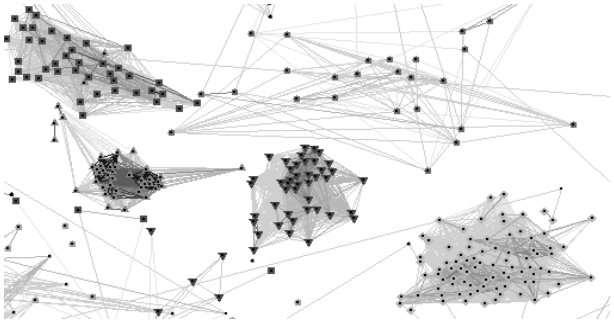


Figure 1: Typical section of CLANS clustering in the His-phosphatase/mutase superfamily.

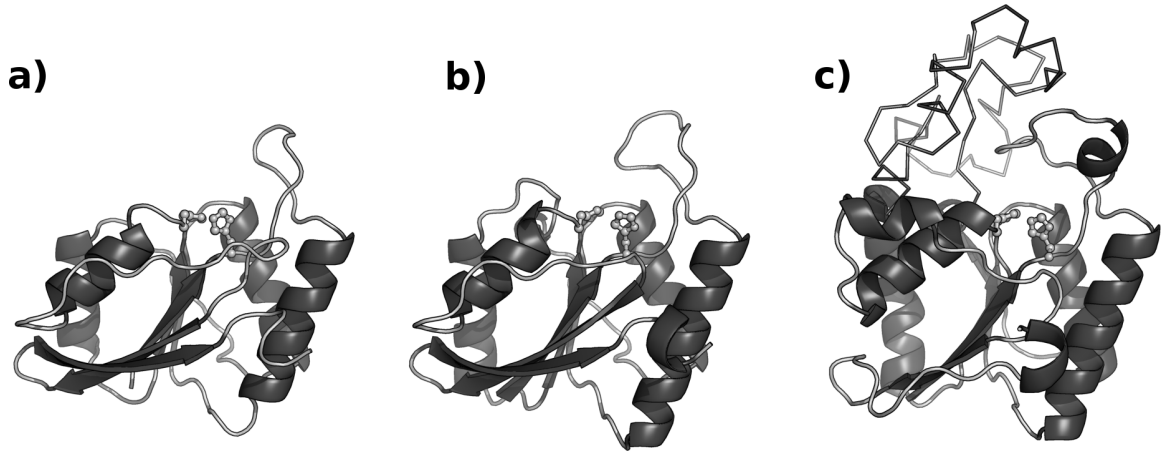


Figure 2: Comparison of (a) a model of a representative of group 20 with crystal structures of (b) SixA and (c) dPGM. Group 20, like SixA, lacks the typical domain insert drawn as Cα trace in (c). Catalytic histidine residues are drawn as ball-and-stick for orientation purposes.

4. REFERENCES

1. Friedberg I. 2006. Automated protein function prediction--the genomic challenge. *Brief Bioinform* 7:225-242.

2. Frickey T. and Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 20:3702-3704.

3. Marchler-Bauer A., *et al.* 2007. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* 35:D237-40.

4. Rigden D.J., Bagyan I., Lamani E., Setlow P. and Jedrzejas M.J. 2001. A cofactor-dependent phosphoglycerate mutase homolog from Bacillus stearothermophilus is actually a broad specificity phosphatase. *Protein Sci.* 10:1835-1846.

Identification of catalytic residues based on destabilizing properties and an SVM classifier

Ran Yahalom[1], Dan Reshef[2], Nir Kalisman[1], Yan Gleyzer[1], Boaz Lerner[3] and Chen Keasar[1,2],*

[1]Department of Computer Science, [2]Department of Life Sciences and [3]Department of Electrical and Computer Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel

*To whom correspondence should be addressed: Keasar@cs.bgu.ac.il

1. INTRODUCTION

Machine learning (ML) offers a generic scheme for the identification of functional residues in structural genomics (SG) targets (1,2). It represents each residue by a vector of features and provides a transformation from the feature space to the binary functional/non-functional space. The current work suggests a set of structure-based features that takes advantage of two characteristics of catalytic residues: their destabilizing nature (3,4) and their frequent burial close to the protein's center of mass (5).

To estimate the destabilizing role of residues, we use eight energy terms from the MESHI molecular modeling package (6). A similar approach was applied in previous studies (7,8) that focused on the electrostatic interactions. Here, we test a wider range of energy terms and improve their individual predictive power by spatial summation. We use the SVM (9) ML technique in order to identify residues based on these energy terms.

2. METHODS

The enzymes and residue annotations used in this study were extracted from the Catalytic Site Atlas (CSA, version 2.0.4) (10). In an attempt to mimic the SG scenario, we excluded hetero-oligomers, mutants and structures with heteroatoms less than 5Å away from a catalytic residue from the data set. The remaining structures were further filtered for high resolution (up to 2.5Å) and non-redundancy (less than 30% pairwise sequence identity). The final data set includes 166 enzyme structures, consisting of 577 catalytic residues and 55,115 non-catalytic residues.

Eight energy terms were tested for their predictive potential: electrostatics, solvation, torsion angle probabilities and propensities, bond, angle out-of-plane and distance-form-centroid (DFC). These are atom-based terms and applying them to the protein atoms results in eight values ε_i^k for each atom $i, k \in \{1,...,8\}$. Each atom is then tagged with spatial summations $E_i^k = \sum_j \varepsilon_j^k e^{-\alpha d_{ij}^2}$ of these values, where j runs over all of the protein's atoms, d_{ij} is the distance between the i^{th} and j^{th} atoms, α is 1 for distance-from-centroid, 0.05 for propensity and 0.01 for all the other terms. The feature vector of each residue r is $\left(Pr, E_r^1,...,E_r^8\right)$, where Pr is 0 or 1 for hydrophobic or polar residues respectively, and $E_r^k = avg(E_i^k \mid i \in \{atoms \text{ of } r\})$ where k is one of the eight terms.

Three SVM classifiers based on linear, polynomial and radial basis function (RBF) kernel types were tested. The classifiers were optimized over a grid of parameters. The performance of each SVM was evaluated using a 10-fold cross-validation experiment.

Statistical significance of energy distribution differences was calculated using the Kolmagorov-Smirnov two-sampled one-tailed test and the Wilcoxon two-sample two-tailed ranksum test. Performance of classifiers was estimated using the Mathews Correlation Coefficient (MCC) and the area under the receiver operating characteristic (AUC) curve.

3. RESULTS

We first compared the energy distributions of catalytic and non-catalytic residues for each of the eight energy terms. The catalytic residues scored significantly higher energy values ($P<10^{-3}$) for seven of the eight terms (all but out-of-plane). This result is even more pronounced when the energies are replaced by their spatial summations, as demonstrated in Figure 1 for the solvation term.

Though statistically significant compared with a random predictor, the ability of the individual terms to predict catalytic residues is not strong. To gain better predictions, we trained SVM classifiers on the energy terms. The best SVM (polynomial kernel) provides an MCC value of 0.737 and an AUC value of 0.933 (Figure 2), which are comparable to those of SVMs that use evolutionary information (1,2).

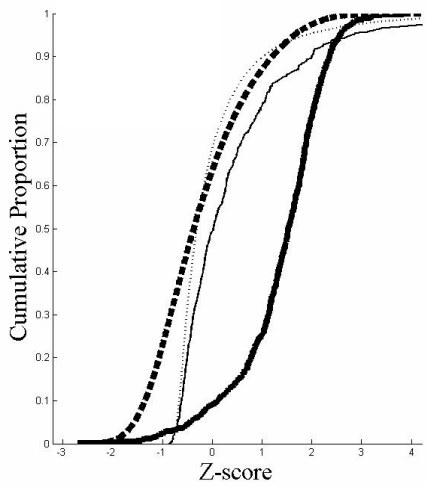


Figure 1. Solvate energy distributions of catalytic (solid lines) vs. non-catalytic (dashed lines) residues. The thin lines represent solvation energies. Thick lines represent spatial summation of these energies. The energy values were transformed to Z-score in order to accommodate the magnitude differences between the energies and their sums. The separation between the catalytic and non-catalytic curves increases considerably after the spatial summation (thick lines), implying stronger predictive power.

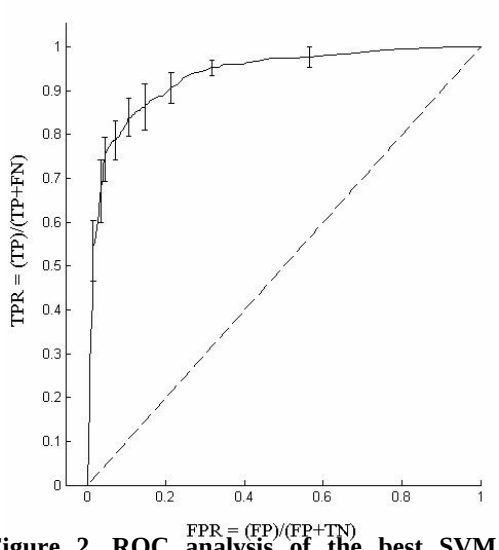


Figure 2. ROC analysis of the best SVM classifier (polynomial kernel). Vertical error bars represent ±1 standard deviation of the TPR (true positive rate) in a ten-fold cross validation experiment. The dashed line represents a random predictor.

4. REFERENCES

1. Petrova NV, Wu CH 2006. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics* **7**:312.

2. Youn E, Peters B, Radivojac P, Mooney SD 2007. Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci* 16(2):216-226

3. Warshel A. 1978. Energetics of enzyme catalysis. *Proc Natl Acad Sci U S A* 75:5250-5254.

4. Schreiber, G., Buckle, A.M. and Fersht, A.R. 1994. Stability and function: two constraints in the evolution of barstar and other proteins. *Structure* 2:945-951

5. Ben-Shimon, A. and Eisenstein, M. 2005. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. *J Mol Biol* 351:309-326

6. Kalisman N., Levi A., Maximova T. Reshef D., Zafriri-Lynn S., Gleyzer Y. and Keasar C. 2005. MESHI: a new library of Java classes for molecular modeling. *Bioinformatics* 21:3931-3932

7. Elcock, A.H. 2001. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J Mol Biol*, 312:885-896

8. Bate, P. and Warwicker, J. (2004) Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods, J Mol Biol, 340:263-276.

9. Platt J.C. 1998. Sequential minimal optimization: a fast algorithm for training support vector machines. *Microsoft Research Tech Report* (MSR-TR-98-14)

10. Porter, C.T., Bartlett, G.J. and Thornton, J.M. 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, 32, D129-133