Journal of Bioinformatics and Computational Biology
Vol. 11, No. 1 (2013) 1340008 (13 pages)

DOI: 10.1142/S0219720013400088

Imperial College Press
www.icpress.co.uk

MOTIF DISCOVERY WITH DATA MINING IN 3D PROTEIN STRUCTURE DATABASES: DISCOVERY, VALIDATION AND PREDICTION OF THE U-SHAPE ZINC BINDING ("HUF-ZINC") MOTIF

SEBASTIAN MAURER-STROH

Bioinformatics Institute (BII)
*Agency for Science and Technology (A*STAR)*
30 Biopolis Street, #07-01, Matrix, 138671, Singapore

School of Biological Sciences (SBS), Nanyang Technological University (NTU)
60 Nanyang Drive, 637551, Singapore
sebastianms@bii.a-star.edu.sg

HE GAO

Bioinformatics Institute (BII)
*Agency for Science and Technology (A*STAR)*
30 Biopolis Street, #07-01, Matrix, 138671, Singapore

NUS Graduate School for Integrative Sciences and Engineering
National University of Singapore
Centre for Life Sciences, #05-01, 28 Medical Drive, Singapore 117456
gaohe@nus.edu.sg

HAO HAN

Bioinformatics Institute (BII)
*Agency for Science and Technology (A*STAR)*
30 Biopolis Street, #07-01, Matrix, 138671, Singapore
hanh@bii.a-star.edu.sg

LIES BAETEN

VIB Switch Laboratory, Katholieke Universiteit Leuven
Herestraat 49, Box 802, 3000 Leuven, Belgium
lies.baeten@gmail.com

JOOST SCHYMKOWITZ

VIB Switch Laboratory, Katholieke Universiteit Leuven
Herestraat 49, Box 802, 3000 Leuven, Belgium
joost.schymkowitz@switch.vib-kuleuven.be

FREDERIC ROUSSEAU

VIB Switch Laboratory, Katholieke Universiteit Leuven
Herestraat 49, Box 802, 3000 Leuven, Belgium
frederic.rousseau@switch.vib-kuleuven.be

LOUXIN ZHANG
Department of Mathematics, National University of Singapore
10 Lower Kent Ridge Road, Singapore 119076
matzlx@nus.edu.sg

FRANK EISENHABER
Bioinformatics Institute (BII)
*Agency for Science and Technology (A*STAR)*
30 Biopolis Street, #07-01, Matrix, 138671, Singapore
and
Department of Biological Sciences (DBS)
National University of Singapore (NUS)
8 Medical Drive 4, 117597, Singapore
and
School of Computer Engineering (SCE)
Nanyang Technological University (NTU)
50 Nanyang Drive, 637553, Singapore
franke@bii.a-star.edu.sg

Received 1 October 2012
Revised 25 October 2012
Accepted 4 December 2012
Published 10 January 2013

Data mining in protein databases, derivatives from more fundamental protein 3D structure and sequence databases, has considerable unearthed potential for the discovery of sequence motif—structural motif—function relationships as the finding of the U-shape (Huf-Zinc) motif, originally a small student's project, exemplifies. The metal ion zinc is critically involved in universal biological processes, ranging from protein-DNA complexes and transcription regulation to enzymatic catalysis and metabolic pathways. Proteins have evolved a series of motifs to specifically recognize and bind zinc ions. Many of these, so called zinc fingers, are structurally independent globular domains with discontinuous binding motifs made up of residues mostly far apart in sequence. Through a systematic approach starting from the BRIX structure fragment database, we discovered that there exists another predictable subset of zinc-binding motifs that not only have a conserved continuous sequence pattern but also share a characteristic local conformation, despite being included in totally different overall folds. While this does not allow general prediction of all Zn binding motifs, a HMM-based web server, Huf-Zinc, is available for prediction of these novel, as well as conventional, zinc finger motifs in protein sequences. The Huf-Zinc webserver can be freely accessed through this URL (http:// mendel.bii.a-star. edu.sg/ METHODS/hufzinc/).

Keywords: Datamining, protein structure database, protein structural motif, zinc binding, zinc finger, HMM, protein sequence motifs.

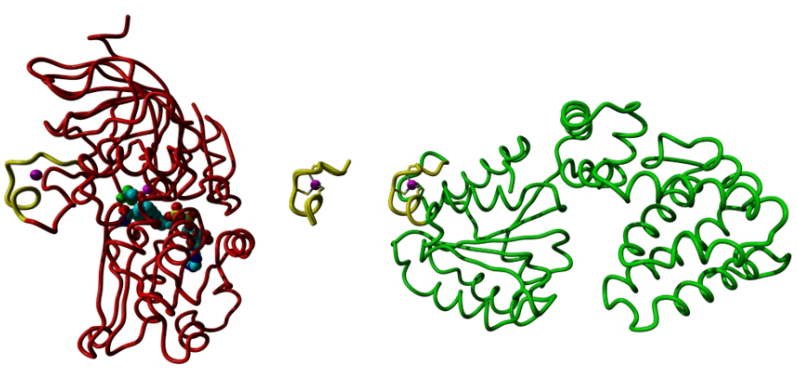
1. Introduction

An increasingly larger number of databases derived from basic collections such as PDB[1] and UniProt[2] provide the information in a preprocessed, more convenient for script-guided searches form and these databases and associated tools enable new, by far not exhausted possibilities for discoveries of protein sequence-structure-function relationships by data mining. Such searches are greatly supported by the myriad of

WWW servers and command line tools that are provided by the community. These projects can, at the beginning, be quite small and provide excellent opportunities for students entering the field to find interesting new insights without spending years of work. In this report, we describe the story of discovering the Huf-Zinc (U-turn zinc-binding) structural motif common to proteins with diverse structures that started with a hypothesis-free mining of occurrences of conserved short motifs in the BriX database.[3,4]

Metal-binding structural motifs are of great importance for sequence-based function prediction and many of them still await description. Zinc is known to play important roles in many biological processes, which has been extensively studied.[5] Thus, accurate prediction of Zinc binding sites, therefore, is greatly supported by the knowledge of the respective zinc-binding motifs. Most published methods rely on protein structure information for the prediction which limits them to a small subset of proteins where 3D structures are available.[6,7] Sequence-based methods, on the other hand, suffer from much lower specificity than their structural counterparts which could be due to taking a generalist approach mixing different types of zinc-binding motifs together. Various methods and servers have been developed for predicting zinc binding sites by identifying similarities in sequence features in homologous proteins.[8,9]

While overall or domain-based sequence and structural similarity is commonly used to suggest conservation in function.[11,12] looking into conserved local 3D structures that can be identified based on characteristic short sequence motifs may extend our capabilities for function annotation. Regions sharing not only a local sequence but also structural motif likely have similar functions.[13] As an example, we can see from Fig. 1 that the alcohol dehydrogenase and N-domain of the delta prime subunit from DNA polymerase III share a common local structural motif serving as a



(a) (b)

Fig. 1. Example for a locally similar structure (U-shape zinc binding motif) in the context of different overall folds. (a) Alcohol dehydrogenase (PDB:1a71, red); (b) delta prime subunit of DNA polymerase III N-domain (PDB:1a5t, green); Center: zinc ion (magenta) and superimposed zinc binding motif (yellow). This figure was created using Yasara.[10]

zinc binding site although the overall folds are totally different. Based on this assumption, we have developed a systematic methodology to discover a novel predictable subset of zinc-binding motifs that not only have a conserved sequence pattern but also share a characteristic local conformation, despite being included in totally different overall folds. We use an HMM-based approach to predict these novel, as well as conventional, zinc finger motifs in protein sequences.

2. Results

2.1. *Mining for potentially new structural motifs in the database BriX and selection of interesting motif targets*

As outlined in the workflow in Fig. 2, we first collected a dataset of short structurally conserved motifs by retrieving 1896 classes of 7 residues length with structural clustering threshold of 0.8 Angstrom RMSD from the BriX database.[3,4] We further filtered out classes with less than 10 sequences and ranked the remainder by sequence motif conservation within the gapless sequence alignment of individual classes. The conservation measure used was based on the average sum of BLOSUM62 substitutions between occurring amino acid types weighted by observed frequencies (similar to the Henikoff-style sum of pairs conservation in AL2CO[14]) and by the square root of number of sequences in the alignment. We selected those conserved motif classes whose cumulative sum of conservation values covers 50% of the overall conservation information (sum of all values). This left us an initial dataset with 353 classes.

Next, we searched Prosite[15] and CompariMotif[16] for selecting known motifs in the initial dataset and examined the 3D structure of exemplary sequences in these classes using SwissPDB-Viewer.[17] We found that several conserved classes appear to match to metal-binding motifs and that the respective peptide chains were located in the vicinity of metal ions. At this stage, we decided to focus on conserved motif classes that have cysteines and/or aspartates, amino acid residues that can interact with metal ions, in their sequences and that have metal ions in close proximity to the peptide chain segment in the respective structures.

In the case of the metal-binding motif CxxC, the local 3D structure showed that the two critical cysteines in CxxC are typically followed by a third cysteine two residues away and, therefore, we extended the motif to CxxCxxC (Fig. 1). This motif has a conserved local structure associated with the conserved sequence pattern. From the 3D structure of metal-binding proteins with this motif, the local conformation of this motif can be better described to be in a "U-shape", "horseshoe"-like ("Huf"=horseshoe in German) form with the three cysteines surrounding the bound metal ion, typically a zinc.

Based on above findings, we compiled datasets of sequences containing the zinc binding motif and also a negative sequence set. First, we manually selected sequences that have at least two cysteines and two of the three important residues interacting with the Zinc using the MSDMotif server.[18] After removing redundancy at 90% sequence identity level, we collected a positive set of Zinc binding sequences. As

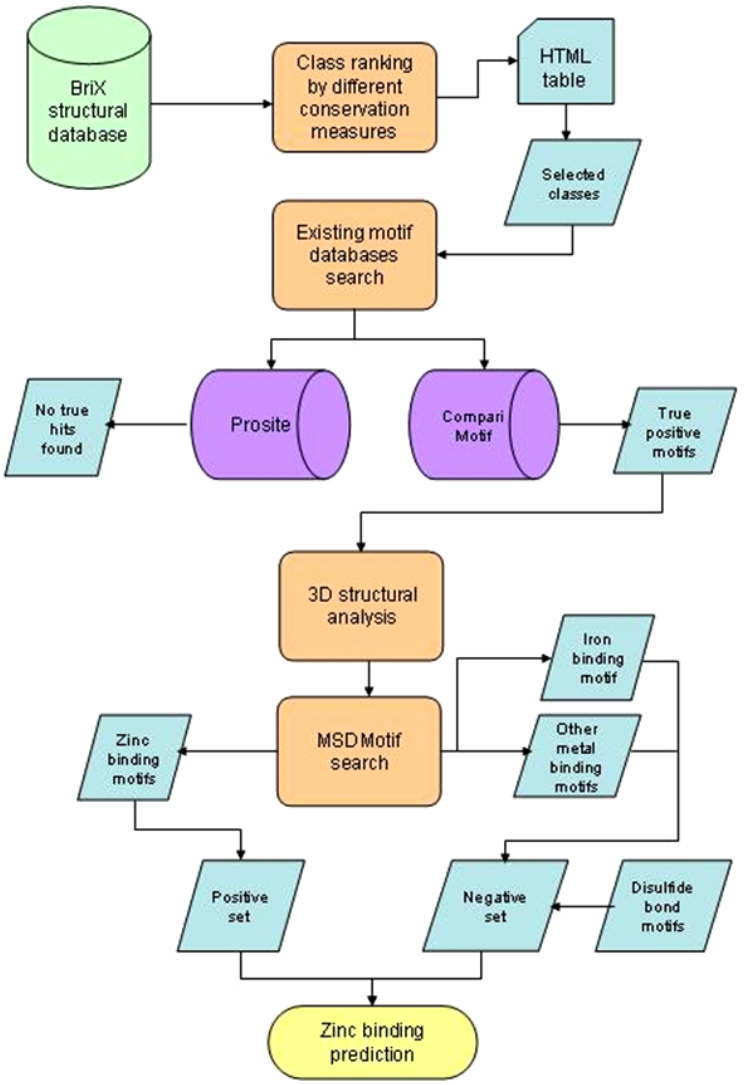


Fig. 2. Outline of the discovery process for the U-shape zinc binding motif. First we narrow down candidates of small structural motifs with sequence conservation. After finding zinc-binding sites with conserved sequence and structural motif, we create positive and negative data sets which were used for developing a predictor for the U-shape zinc-binding motif.

indicated in the sequence logo[19] given in Fig. 3, there can be partial substitutions of cysteines with histidines or negatively charged aspartates or glutamates. Also, the intermittent and adjacent positions appear to have some mild restrictions in preferred amino acid types.

The sequences for the negative data set comprise two parts: (1) sequences from the MSDMotif search result with at least two of the three important residues

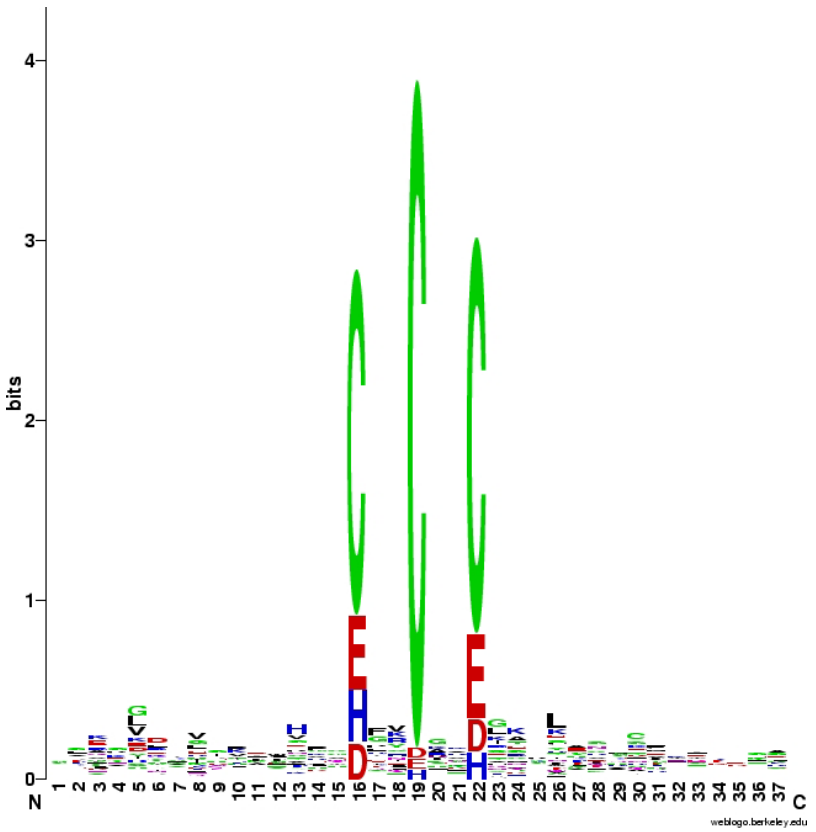


Fig. 3. Sequence logo for the U-Shape Zinc binding motif showing the 3 critical positions with clear preference for cysteines but partially allowing histidines and negative charges. Also the amino acid types at positions inbetween and shortly before and after do not appear random and hence contribute to the specific motif characteristics leading to a conserved local structure as shown in Fig. 1.

interacting with iron and other metals. (2) sequences parsed from SwissProt/UniProt[2] for which the central cysteine residue is involved in the formation of disulfide bridges and, hence, cannot bind metals. Similarly, redundancy was removed at 90% sequence identity level and the negative set was formed.

2.2. *Development of a HMM-based predictor for the newly discovered motif*

In order to further distinguish the positive and negative sets to improve the prediction power, the entropy difference between the initial positive and negative sets were studied. The two sets were sent to Shannon Entropy-Two, an online tool that gives the Shannon entropy difference between defined background and query sets by Monte Carlo randomizations.[20] Since there are 192 sequences in the positive set and they have a length of 37 amino acids, 1920 random sequences of the same length were

selected from UniProt to form the background set. Based on the entropy difference analysis, we decided to shorten the motif by removing the flanking residues that are of low entropy. Therefore, the refined positive set and negative set were selected to be 16 amino acids long. After applying redundancy removal at 90% level, the final positive set contained 191 sequences and the negative set contained 247 sequences. Among the negative sequences, 110 are sequences that bind to iron and other metals, 137 are disulfide bond containing sequences.

Making use of this new data set, we developed a zinc-binding predictor based on a profile Hidden Markov Model of the alignment of known zinc-binding sequences, as implemented in HMMER.[21] Five-fold cross validation was adopted. The whole positive set was randomly divided into five subsets and four of them were used as training set to build an HMM model. The model is used to predict the positive test set which is the remaining subset from the positive set and the negative test set which is the whole negative set. A ROC curve given in Fig. 4 shows the comparison of performance of Huf-Zinc and PredZinc.[9] It can be seen that the true positive rate (TPR) which is equivalent to sensitivity of the 5-fold cross-validated model is up to 70% better than PredZinc at relevant levels of high specificity. For example, at a false positive rate of only 5%, around 85% of the known positive examples were identified. A maximum MCC (Matthews correlation coefficient) value of 0.77 was achieved which still indicates further room for improvements. The noncross-validated model which would be expected to over-predict is also shown for comparison purposes.

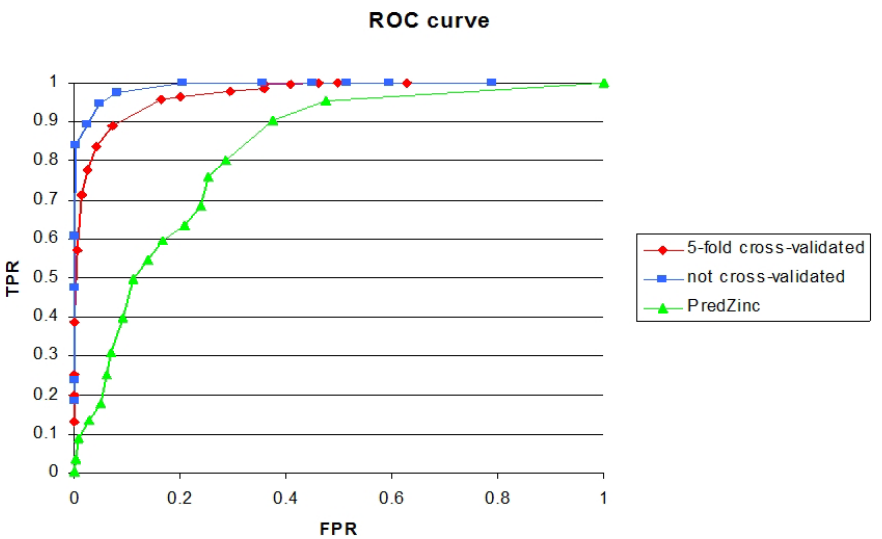


Fig. 4. ROC performance comparison over known U-shape zinc-binding motifs with 3D structure evidence versus negative examples represented by similar sequence motifs binding other metal ions or where the central cysteine forms a disulfide bridge.

Since our datasets were formed by searching the MSDMotif database with motif patterns that were derived from structurally conserved classes and the zinc-binding local structure has been proven by 3D analysis to consist of three cysteines forming a "U-shape", our model specifically predicts zinc-binding motifs that are in that conformation, whereas PredZinc tends to treat all zinc-binding motifs in the same way and does not differentiate different structures. As a result, our model can more precisely predict sequences in the testing set compared to PredZinc. There is an advantage for being specific, as zinc-binding motifs that have different 3D structures can have entirely different properties and are not easily comparable to each other. At the same time, higher specificity also means that only the U-shape motif can be identified while there are also several other, especially discontinuous zinc binding motifs which cannot be detected with our method. However, to our surprise there are also some curious cases where motifs with 3 cysteines arranged as CxxCxxC can be part of multiple structural sites that come together in 3D space (for example PDB entries 2XIG and 2XJY). For all cases with available 3D structure, it remains important to critically compare the predicted sequence motif with the structural context to confirm which residues bind zinc as part of which motif.

A web server named Huf-Zinc is available online at.[22] Users are able to pass their protein sequence to be studied by either pasting it into the sequence input window or by directly uploading it from their local machine. Four options are provided which allow users to opt for high sensitivity (at a bit score cutoff value of 0 which corresponds to 98% sensitivity level), balanced performance (score cutoff 2, corresponds to highest MCC), high specificity (at cutoff value 6 which corresponds to 5% false positive rate or equivalently 95% specificity level) or a customized cutoff. Prediction results are returned in a format showing the sequence predicted as zinc-binding, the start and endg position in the input sequence as well as the e-value. For the convenience of the reader, the sequence sets used for deriving the Huf-Zinc method, the respective HMM and several other materials are also made available at the WWW site.[22]

3. Discussion

The U-shape zinc binding motif identified in this study can be frequently seen also as part of classical zinc finger domains that typically add a fourth cysteine to the U-shape motif through loops of varying lengths and structure. However, more importantly it can also occur in totally different overall folds as exemplified by alcohol dehydrogenase and the N-domain of the delta prime subunit from DNA polymerase III (Fig. 1). To gauge the extent of such motifs that are not found as classical zinc finger domains, we searched the PFAM database[23] with the keyword "zinc finger" and finally selected 68 PFAM motifs through manual curation by confirming zinc-binding for example through an existing structure with bound zinc.

Using the threshold of best overall performance (highest MCC), we used our U-shape zinc binding predictor to search against the UniRef90.[2] Among 2475

proteins predicted by Huf-Zinc, there are 582 (23.5%) that are not predicted by the PFAM zinc finger domain set (despite an optimistic E-value cutoff of 0.1 for the HMMER search with the PFAM domains; see complete list of PFAM domains at.[22]) This suggests that the U-shape motif predictor presented here allows for a substantial increase of zinc binding motifs that can be identified from protein sequences.

It is known that metal atoms play an important role in the structure and stability of proteins and many proteins need to bind one or more metal ions in order to perform their functions. Besides stabilizing protein tertiary/quaternary structures, metal ions are also involved in catalytic mechanism. Therefore, identification/prediction of metal binding sites greatly helps the investigation of the function of experimentally uncharacterized genes and proteins,[24,25] the most challenging task in the post-genomic era.[26] Prediction tools like Huf-Zinc unfold their full value in sequence-analytic environments such as the ANNNOTATOR where they are silently invoked together with dozens of other predictors for any query sequence.[27,28]

Similar characterization for other sequence-function relationships will be useful in the future. The same approach as outlined here can be used to validate and predict further functionally important structural motifs. For example, we encountered the motif DxDxD in our search BriX database search. This motif contains conserved aspartic acid residues with the potential to bind calcium or magnesium.

4. Conclusions

The findings in this study are based on a discovery-based approach integrating sequential and structural information. We identified a novel U-shape zinc-binding motif and the unique sequence and local structure conservation may allow prediction of this specific subset of zinc binding motifs.

Acknowledgment

The authors acknowledge A*STAR for funding this research.

References

1. Kirchmair J, Markt P, Distinto S, Schuster D, Spitzer GM, Liedl KR, Langer T, Wolber G, The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery, *J Med Chem* **51**:7021−7040, 2008.
2. The Universal Protein Resource (UniProt) in 2010, *Nucleic Acids Res* **38**:D142−D148, 2010.
3. Baeten L, Reumers J, Tur V, Stricher F, Lenaerts T, Serrano L, Rousseau F, Schymkowitz J, Reconstruction of protein backbones from the BriX collection of canonical protein fragments, *PLoS Comput Biol* **4**:e1000083, 2008.
4. Baeten L, Vanhee P, Schymkowitz J, and Rousseau F.2012 Brix Database [Online]. http://brix.switchlab.org/.
5. Stefanidou M, Maravelias C, Dona A, Spiliopoulou C, Zinc: A multipurpose trace element, *Arch Toxicol* **80**:1−9, 2006.

6. Ebert JC, Altman RB, Robust recognition of zinc binding sites in proteins, *Protein Sci* **17**:54−65, 2008.
7. Zhao W, Xu M, Liang Z, Ding B, Niu L, Liu H, Teng M, Structure-based de novo prediction of zinc-binding sites in proteins of unknown function, *Bioinformatics* **27**:1262−1268, 2011.
8. Passerini A, Andreini C, Menchetti S, Rosato A, Frasconi P, Predicting zinc binding at the proteome level, *BMC Bioinformatics* **8**:39, 2007.
9. Shu N, Zhou T, Hovmoller S, Prediction of zinc-binding sites in proteins from sequence, *Bioinformatics* **24**:775−782, 2008.
10. Krieger E, Koraimann G, Vriend G, Increasing the precision of comparative models with YASARA NOVA−a self-parameterizing force field, *Proteins* **47**:393−402, 2002.
11. Punta M, Ofran Y, Erratum, *PLoS Comput Biol* **4**:2008.
12. Punta M, Ofran Y, The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function, *PLoS Comput Biol* **4**: e1000160, 2008.
13. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen L, Bridging protein local structures and protein functions, *Amino Acids* **35**:627−650, 2008.
14. Pei J, Grishin NV, AL2CO: Calculation of positional conservation in a protein sequence alignment, *Bioinformatics* **17**:700−712, 2001.
15. Sigrist CJ, Cerutti L, de CE, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N, PROSITE, a protein domain database for functional characterization and annotation, *Nucleic Acids Res* **38**:D161−D166, 2010.
16. Edwards RJ, Davey NE, Shields DC, CompariMotif: Quick and easy comparisons of sequence motifs, *Bioinformatics* **24**:1307−1309, 2008.
17. Guex N, Peitsch MC, SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling, *Electrophoresis* **18**:2714−2723, 1997.
18. Golovin A, Henrick K, MSDmotif: Exploring protein sites and motifs, *BMC Bioinformatics* **9**:312, 2008.
19. Crooks GE, Hon G, Chandonia JM, Brenner SE, WebLogo: A sequence logo generator, *Genome Res* **14**:1188−1190, 2004.
20. Los Alamos National Laboratory.2012 Shannon Entropy-Two [Online]. http://hcv.lanl.gov/content/sequence/ENTROPY/entropy.html.
21. Eddy SR, Profile hidden Markov models, *Bioinformatics* **14**:755−763, 1998.
22. 2012 Huf-Zinc WWW server [Online]. http://mendel.bii.a-star.edu.sg/METHODS/hufzinc/.
23. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A, The Pfam protein families database, *Nucleic Acids Res* **38**:D211−D222, 2010.
24. Andreini C, Banci L, Bertini I, Rosato A, Zinc through the three domains of life, *J Proteome Res* **5**:3173−3178, 2006.
25. Andreini C, Banci L, Bertini I, Rosato A, Counting the zinc-proteins encoded in the human genome, *J Proteome Res* **5**:196−201, 2006.
26. Eisenhaber F, A decade after the first full human genome sequencing: When will we understand our own genome?, *J Bioinform Comput Biol* **10**:1271001, 2012.
27. Schneider G, Wildpaner M, Sirota FL, Maurer-Stroh S, Eisenhaber B and Eisenhaber F. Integrated Tools for Biomolecular Sequence-Based Function Prediction as Exemplified by the ANNOTATOR Software Environment. In: Data Mining Techniques for the Life Sciences, edited by Carugo O and Eisenhaber FE. New York: Humana Press and Springer Business Media, 2009, p. 257−268.

28. Schneider G, Sherman W, Kuchibhatla D, Ooi HS, Sirota FL, Maurer-Stroh S, Eisenhaber B and Eisenhaber F. Protein sequence-structure-function-network links discovered with the ANNOTATOR software suite: Application to Elys/Mel-28. In: Computational Medicine, edited by Trajanoski Z. Vienna: Springer, 2012, p. 111−143.

Sebastian Maurer-Stroh studied theoretical biochemistry at the University of Vienna and wrote his master and Ph.D. thesis at the renowned Institute of Molecular Pathology (IMP). After the honor of a FEBS and a Marie Curie fellowship at the VIB-SWITCH lab in Brussels, he leads a group of experts in protein sequence analysis as principal investigator in the A*STAR Bioinformatics Institute (BII) since November 2007. He has contributed widely used predictors for posttranslational lipid modifications, amyloid fibre formation and catalyzed new biomedical insights by sequence-based function predictions. He has been at the forefront of research during the 2009 H1N1 pandemic collaborating with hospitals and health authorities in Singapore, Mexico, Brazil and Australia. He initiated a cross-division programme for Human Infectious Diseases at the Bioinformatics Institute that builds upon the expertise of several groups from different backgrounds.

He Gao was an undergraduate student in the Computational Biology Programme, National University of Singapore (NUS) and this work was part of her B.Sc. Honors year project. She is currently a Ph.D. candidate in molecular epidemiology under NUS Graduate school for Integrative Sciences and Engineering.

Hao Han's research area lies in developing biological databases as well as bioinformatics tools for protein function prediction. Han Hao obtained his master at the National University of Singapore. He joined the Bioinformatics Institute as a bioinformatics specialist since November 2006.

Lies Baeten obtained her Ph.D. in the Switch lab in 2010 entitled "Reconstruction of protein structures from polypeptide fragment libraries". She developed the BriX protein fragment database, the LoopBriX loop database and the loop reconstruction algorithm LoopX. She is now an IT consultant.

Joost Schymkowitz's research interest is focused on cellular processes regulated by protein conformational switching. More specifically he is interested in understanding the mechanisms leading to toxic protein aggregation as observed in several important neurological disorders such as Parkinson's, Alzheimer's and Creutzfeld-Jacob diseases. His approach includes the use of home-made structural bioinformatic tools, biophysical analysis as well as cellular biology. He received his Ph.D. in protein engineering from the University of Cambridge followed by post-doctoral research in bioinformatics and protein aggregation at EMBL Heidelberg. Since 2003 he is one of the two group leaders of the Switch laboratory, an independent research unit of the Flemish Interuniversity Institute for Biotechnology (VIB).

Frederic Rousseau's research is focused on understanding the mechanisms gearing protein folding and misfolding and their relation to human disease. In particular he is investigating how protein aggregation affects the interactome by suppressing native interactions but also by introducing novel aggregation-specific interactions. The latter are especially relevant as they are usually associated to gain of function activities such as neurotoxicity (neurodegeneration) or cell proliferation (cancer). Frederic received his Ph.D. in chemistry from the University of Cambridge. After post-doctoral research with Luis Serrano at EMBL Heidelberg, he is one of the two group leaders of the Switch laboratory of the Flemish Interuniversity Institute for Biotechnology (VIB) since 2003.

Louxin Zhang's research interest lies at the interface of mathematics and computer science, genomics, and molecular evolutionary biology. This includes elucidating the mechanism that confers power to spaced seeds for homology search, phylogenetic analysis and evolution of duplicated gene clusters, as well as developing computational methods for integrating comparative genomics data to elucidate cross-species differences and within-species variation and their associations with disease. Zhang Louxin is an associate professor of the National University of Singapore since July 2004.

Frank Eisenhaber's research interest is focused on the discovery of new biomolecular mechanisms with theoretical and biochemical approaches and the functional characterization of yet uncharacterized genes and pathways. Frank Eisenhaber is one of the scientists credited with the discovery of the SET domain methyltransferases, ATGL, kleisins, many new protein domain functions and with the development of accurate prediction tools for post-translational modifications and subcellular localizations. Frank Eisenhaber studied mathematics at the Humboldt-University in Berlin and biophysics and medicine at the Pirogov Medical University in Moscow. He received the Ph.D. from the Engelhardt Institute of Molecular Biology in Moscow (1988). After postdoctoral work at the Institute of Molecular Biology in Berlin-Buch and at the EMBL in Heidelberg, he lead the bioinformatics research group and the IT department at the Institute of Molecular Pathology (IMP) in Vienna (1999−2007). Since August 2007, he is the Director of the Bioinformatics Institute A*STAR Singapore.