Final Report
DE-FG02-96ER62268
Combined Approach to the Inverse Protein Folding Problem
Ruben A. Abagyan – P.I.

have no objection from a patent
standpoint to the publication or
dissemination of this material.
MPD WTS
Office of Intellectual
Property Counsel
DOE Field Office, Chicago
6/12/00
Date

The main scientific contribution of the project "Combined approach to the inverse protein folding problem" submitted in 1996 and funded by the Department of Energy in 1997 is the formulation and development of the idea of the multilink recognition method for identification of functional and structural homologues of newly discovered genes. This idea became very popular after we first announced it and used it in prediction of the threading targets for the CASP2 competition (Critical Assessment of Structure Prediction).

The premise of the idea:

A direct comparison of a new query sequence and a database sequence with known structure or function has only a limited scope of recognition, however, there are large numbers of sequences with unknown structures or functions that can be used to enhance recognition. To achieve our goal, for the first iteration, we built a 'family' or query sequence in the entire space of characterized and uncharacterized sequences. There are now two strategies. The first strategy was later implemented in PSI-BLASE: take the family coming out from the first iteration, convert it into a profile and resubmit the profile. The second strategy (the one which was actually proposed) was to resubmit the search with each of the individual sequences found after the first iteration. These iterations can continue until convergence.

Since the multilink method for sensitive fold recognition was based on a series of pairwise sequence comparisons it required better statistics, as well as the application of the method to real biological problems [1,2,3].

Optimization of sequence comparison aimed at structural recognition

First, we compiled a database of protein domain sequences that were significantly similar to the sequences of all the protein domains with known 3D structure according to the SCOP classification. To develop a sensitive sequence-based procedure for fold recognition, we performed an exhaustive global alignment (with zero end gap penalties) between sequences of protein domains with known 3D folds. The subset of 1.3 million alignments between sequences of structurally unrelated domains was used to derive a set of analytical functions which represented the probability of structural significance for any sequence alignment at a given sequence identity, sequence similarity and alignment score. Analysis of the overlap between structurally significant and insignificant alignments showed that sequence identity and sequence similarity measures are poor indicators of structural relatedness in the "twilight zone", while the alignment score allows much better discrimination between alignments of structurally related and unrelated sequences for a wide variety of alignment settings. A fold recognition benchmark was used to compare eight different substitution matrices with eight sets of gap penalties. The best performing matrices were Gonnet and Blosum50 with normalized

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

gap penalties of 2.4/0.15 and 2.0/0.15, respectively, while the positive matrices were the worst performers. The derived functions and parameters can be used for fold recognition [1].

Application of the method for recognition of remote homologues

The optimized fold recognition protocol led to an important discovery. The newly derived statistics of structural significance of the semi-global sequence alignment scores, as well as the optimized parameters of sequence comparison allowed us to establish a reliable structural similarity between the recently identified gene TSG101 and a family of ubiquitin-conjugating enzymes (UBC) [2]. The TSG101 gene contains deletions in 40% of analyzed primary breast cancers but not in normal breast tissue from the same patients, suggesting that it may be important in the development of non-hereditary breast cancer, which accounts for 95% of the disease. The top hits in the SWISS-prot database were the yeast enzyme (YCL008c) with the p-value of statistical significance as low as 10^{-16} and the ubiquitin conjugating enzyme UBC2_YEAST (p-value 10^{-7}). Further support for the structural relationship between TSG101 and ubiquitin-conjugating enzymes was provided by the analysis of a 3D model of TSG101, which was constructed and energy minimized using the ICM method. The model shows that all the essential residues surrounding the active site cysteine and conserved in UBCs are also present in TSG101 [2]. A PH-domain study was another extensive application of the method to an important biological problem [10].

CED-4, the key component of a programmed cell death pathway, is one of the most investigated proteins currently in biology, yet no functional or structural information has been derived from its sequence. The multilink principle and the newest structural statistics led us to the discovery of a structural relationship between CED4 and the AAA family of ATPases. As foreseen in our proposal, the direct sequence comparison does not identify any reliable similarity. However, an intermediate genomic sequence for *C. elegans* establishes the necessary link between CED4 and the AAA family. The probabilities of the two links are the following: CED4- $\rightarrow$ C.eleg.protein (p-value= $10^{-4.45}$), and C.eleg.protein $\rightarrow$ AAA ATPase (p-value= $10^{-19.4}$). A careful analysis of the mutations and motifs confirmed the finding.

CAD

To compare the results of the inverse protein folding we developed a new measure of structural similarity [5]. This unified measure attempts to evaluate the accuracy of three-dimensional atomic protein models. A normalized sum of absolute differences of residue-residue contact surface areas are calculated for a reference structure and a model. It employs a more rigorous quantitative evaluation of a contact than previously proposed contact count based measures. We argue that the contact area difference (CAD) number is a robust single measure to evaluate protein structure predictions in a wide range of model accuracies, from ab initio and threading models to models by homology, since it reflects both the local secondary structure and packing geometry, is smooth, continuous and threshold-free, is not sensitive to typical crystallographic errors and ambiguities,

adequately penalizes domain and/or secondary structure rearrangements and protein plasticity, and has consistent linear and matrix representations for more detailed analysis. The CAD quality of crystallographic structures, NMR structures, models by homology, and unfolded and misfolded structures is evaluated. It was shown that the CAD number discriminates between models better than the Cartesian root-mean-square deviation (cRMSD). The source code of the program calculating the CAD measures was made available for the scientific community and was used by the assessors of the CASP competition.

Modeling

Structural recognition enables one to build models and use them to reach biological conclusions. We used the optimized matrices and alignment parameters to derive models of several biomolecules in order to understand their function and structure. In a series of three papers [8,9,13] we showed how we built three-dimensional models and identified the active sites in CD59 proteins. Our analysis enabled us to understand the function of this class of proteins and explained the complex pattern of cross specificity observed in different species [9,13]. A number of models were built (e.g. 12) based on our optimal parameters.

The research conducted over the life of this project has prepared us to meet the challenges of future human genome projects.

Publications relevant to the grant

- [1]. Abagyan, R.A. & Batalov, S. Do aligned sequences share the same fold? *J. Mol. Biol.*, 273: 355-368, 1997.
- [2]. Cardozo, T., and Abagyan, R. Molecular Modeling of the Domain Shared Between CED-4 and its Mammalian Homologue Apaf-1: A Structural Relationship to the G-proteins. *J. Mol. Model.*, 4: 83-93, 1998.
- [3]. Koonin, E.V. & Abagyan, R.A. TSG101 may be the prototype of a class of dominant negative ubiquitin regulators *Nature Genetics*, 16: 330-331, 1997.
- [4]. Abagyan, R., Batalov, S., Cardozo, T., Totrov, M., Webber, J. and Zhou, Y. Homology Modeling with Internal Coordinate Mechanics: Deformation Zone Mapping and Improvements of Models via Conformational Search. *Proteins: Struct. Funct. & Gen.*, Suppl. 1: 29-37, 1997.
- [5]. Abagyan, R. & Totrov, M. Contact Area Difference (CAD): A robust measure to evaluate accuracy of protein models. *J. Mol. Biol.*, 268: 678-685, 1997.

- [6]. Abagyan, R.A. Protein structure prediction by global energy optimization. Computer simulation of biomolecular systems: Theoretical and experimental applications, Kluwer Academic P363-394: 363-394, 1997.
- [7]. Rashin, A.A., Rashin, B.H., Rashin A., Abagyan, R. Evaluating the energetics of empty cavities and internal mutations in proteins. *Protein Science*, 6: 2143-2158, 1997.
- [8]. Yu, J., Abagyan, R., Dong, S., Gilbert, A., Nusenzweig, V. & Tomlinson, S. Mapping of the Active Site of CD59. *J. Expt. Medicine*, 185: 745-754, 1997.
- [9]. Yu, J., Dong, S., Rushmere, N.K., Morgan, B.P., Abagyan, R. & Tomlinson, S. Mapping the regions of the complement inhibitor CD59 responsible for its species selective activity. *Biochemistry*, 36: 9423-9428, 1997.
- [10]. Isakoff SJ, Cardozo T, Andreev J, Li Z, Ferguson KM, Abagyan R, Lemmon MA, Aronheim A, Skolnik EY Identification and analysis of PH domain-containing targets of phosphatidylinositol 3-kinase using a novel in vivo assay in yeast. *EMBO J*, 17: 5374-5387, 1998.
- [11]. Mayorov, V.N. and Abagyan, R.A. Energy strain in three-dimensional protein structures *Folding and Design*, 3 : 259-269, 1998.
- [12]. Patel, I.R., Attur M.G., Patel R.N., Stuchin S.A., Abagyan R.A., Abramson S.B., Amin A.R. TNF-alpha convertase enzyme from human arthritis-affected cartilage: isolation of cDNA by differential display, expression of the active enzyme, and regulation of TNF-alpha. *J. Immunol.*, 160: 4570-4579, 1998.
- [13]. Srivastava S, Osten P, Vilim FS, Khatri L, Inman G, States B, Daly C, DeSouza S, Abagyan R, Valtschanoff JG, Weinberg RJ, Ziff EB Novel anchorage of GluR2/3 to the postsynaptic density by the AMPA receptor-binding protein ABP. *Neuron*, 21: 581-591, 1998.
- [14] Zhang HF, Yu J, Chen S, Morgan BP, Abagyan R, Tomlinson S. (1999) Identification of the individual residues that determine human CD59 species selective activity. *J Biol Chem* 1999 Apr16;274(16):10969-74.