BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(7)

HORIBALFRE program: Higher Order Residue Interactions Based ALgorithm for Fold REcognition

Pandurangan Sundaramurthy [1, 2], Raashi Sreenivasan [1, 3, 4#], Khader Shameer [1,5#], Sunita Gakkhar [2] & Ramanathan Sowdhamini [1*]

[1]National Center for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore - 560065, India; [2]Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee -247667, India; [3]Centre for Biotechnology, Anna University, Chennai - 600025, India; [4]University of Wisconsin-Madison, Madison, WI 53706-1481, USA; [5]Division of Cardiovascular Diseases, Mayo Clinic, Rochester, MN 55901 USA; Ramanathan Sowdhamini – Email: mini@ncbs.res.in; *Corresponding author: #Authors contributed equally to this work.

Received October 21, 2011; Accepted November 24, 2011; Published December 10, 2011

Abstract:
Understanding the functional and structural implication of a protein encoded in novel genes using function association or fold recognition approaches remains to be a challenging task in the current era of genomes, metagenomes and personal genomes. In an attempt to enhance potential-based fold-recognition methods in recognizing remote homology between proteins, we propose a new approach “Higher Order Residue Interaction Based ALgorithm for Fold REcognition (HORIBALFRE)”. Higher order residue interactions refer to a class of interactions in protein structures mediated by C_α or C_β atoms within a pre-defined distance cut-off. Higher order residue interactions (pairwise, triplet and quadruplet interactions) play a vital role in attaining the stable conformation of a protein structure. In HORIBALFRE, we incorporated the potential contributions from two body (pairwise) interactions, three body (triplet interactions) and four-body (quadruple interaction) interactions, to implement a new fold recognition algorithm. Core of HORIBALFRE algorithm includes the potentials generated from a library of protein structure derived from manually curated CAMPASS database of structure based sequence alignment. We used Fischer’s dataset, with 68 templates and 56 target sequences, derived from SCOP database and performed one-against-all sequence alignment using T-Coffee. Various potentials were derived using custom scripts and these potentials were incorporated in the HORIBALFRE algorithm. In this manuscript, we report outline of a novel fold recognition algorithm and initial results. Our results show that inclusion of quadruplet class of higher order residue interaction improves fold recognition.

Keywords: fold recognition, fold prediction algorithm, protein folding, residue interactions, higher order residue interactions.

Background:

Protein sequence encodes the fundamental structural unit of life: “protein structure” and protein structure defines its biological function **[1-5]**. Knowledge of the structure and function of proteins is, therefore, important in the area of biomedical sciences. Protein structure prediction from sequence information has been a grand challenging problem in molecular biology for the last forty years. Nature conserves structure core, due to convergent evolution, and the number of unique structural (domain) folds in nature is possibly limited. The probability for a protein sequence to have a native-like structural fold in Protein Data Bank (PDB) **[6]** is estimated to be 60-70%. Various fold recognition methods based on mathematical, statistical, and computational algorithms have been developed to predict possible from sequence information. For example, contact map model-based pseudoenergy function incorporating pairwise residue interaction potential by allowing variable gaps have been developed and implemented to predict protein 3D structure through fold recognition method. The ability of these fold recognition methods to accurately distinguish the correct, folded structure from moderately distorted (misfolded) structures is limited **[7-14].**

In this manuscript, we propose a new method called as Higher Order Residue Interaction Based ALgorithm for Fold REcognition (HORIBALFRE). We incorporated potential contributions not just from one-body and two-body terms, but also from the three-body (triplet interactions) and the four-body (quadruple interaction) interactions, to improve the performance of fold prediction using sequence data. Core of HORIBALFRE includes the potentials generated from a structure library derived from CAMPASS database **[15]** of structure based sequence alignment. We used Fischer's dataset, with 68 templates and 56 target sequences, derived from SCOP database and performed one-against-all sequence alignment using T-Coffee **[16]**. These potentials were incorporated into HORIBALFRE algorithm. Currently, the algorithm was applied to the Fischer's dataset with 68 template-target pairs.

Sequence databases are experiencing an unprecedented growth in the post-genome era due to automated sequencing techniques. Annotation of the sequences by computational approaches using structure and sequence based methods are getting increasingly important **[1-3, 7, 12, 17-22].** As attempts to sequence entire genomes increases the number of protein sequences by a factor of two each year, the gap between sequence and structural information stored in public databases growing rapidly **[23]**. To fill the sequence-structure-function gap and to completely understand the function role of a protein and its multitude of cellular interactions, the knowledge of 3D structures is very crucial. As the cost of sequencing technologies are decreasing at an increased rate, the experimental approaches for high–throughput characterization of protein structures using X-ray crystallography and NMR spectroscopy remains a challenge due to cost and laborious nature and often unsuccessful experimental processes. As an alternative, theoretical and computational methods to predict the structure from sequence such as homology, *ab initio,* and fold recognition are widely employed. Homology (comparative) modeling, attempts to predict protein structure on the strength of a protein sequence similarity to another proteins with known structures. Even though it has been the most reliable technique for protein structure prediction, its dependence on alignment quality and the existence of good homologue, indicate it is not applicable to a large fraction of protein sequences which are not within 'structural distance' in sequence space and only 10% of the sequences are modeled **[12, 13, 17]**. *Ab initio* method encompasses any means of calculating co-ordinates of protein structure for a protein sequence from physical principles. Despite a few recent successes on small proteins and short peptides, this method is still not a practical proposition for predicting protein structure due to limitations in computing power and poor understanding of the biophysical forces driving protein folding. The third category of protein structure prediction, falling somewhere between Homology modeling and *ab initio* prediction, is fold recognition.

Methodology:

Conceptual idea behind fold recognition method came from the estimate that there is an ~70% chance that a newly characterized protein with no obvious common ancestry to proteins with a known structure will in fact turn out to share a common fold with at least one protein of known structure in the database**[7, 24].** The objective of fold recognition approach was that given a sequence and a library of structure templates; discover which fold is best compatible with the given sequence **[3, 9, 18, 19, 25-32]**. If the target protein shares significant sequence similarity to a protein of known 3D structure, the fold recognition problem is trivial – simple sequence comparison will identify the correct fold. Threading based approaches could detect structural similarities that are not accompanied by any detectable sequence similarity, and thus, fold recognition is the protein structure prediction method of choice when (1) the sequence identity to any sequence with a known structure, and (2) one or more structures from the structure library represents the true fold of the sequence. Based on the pseudoenergies derived from the statistical analysis of observed protein structures (knowledge - based approach), existing computational methods for fold recognition can be grouped into two major classes: First class of methods employ residue local environments and do not include residue interaction potentials explicitly **[1-3, 8-11, 13, 32-36].** In this kind of method, the prediction speed is fast, but they were not effective in detecting structural similarities between divergent proteins, and between proteins sharing a common fold through convergent evolution (analogous folds). The reason for these limitations is down to the loss of structural information due to residue interactions **[24]**. Second class of methods includes pairwise residue interaction potentials **[3, 18, 33, 35, 37].** However, pairwise residue interactions cannot capture regularities of protein structure and found statistically inadequate to explain the frequency distribution of residue interactions, and consideration of cooperative interactions of higher order may improve the quality of structure prediction **[11, 20, 33, 34, 38-40].** In this manuscript, we introduce the architecture of a new fold recognition algorithm, HORIBALFRE, which employs higher order residue interaction potentials and an integrated approach that also include local environments of residues. We recently showed that a webserver which can compute higher order residue interactions can be used for in-depth structure analysis **[41].** HORIBALFRE is an extension of HORI server and utilize pre-computed amino acid interaction data derived using higher order residue interaction programs developed for HORI server..

Description of HORIBALFRE algorithm:

HORIBALFRE is a multi-step fold recognition algorithm with 5 major steps. A flow-chart of the algorithm is given in **Figure 1.** The following consecutive steps form the core of HORIBALFRE.

(1) Library of target-template (derived from Fischer's dataset)

Target-template library is sourced from Fischer's benchmark dataset **[21]** comprising of 68 unique probe sequences and 56 unique target structures (PDB identifiers of proteins in Fischer's dataset is provided in supplementary material**)**

(2) Alignment of target sequence to the template sequences

T-Coffee [16] is used for the alignment of the target sequence to the template sequence. T-Coffee is used in Global alignment mode and global-local alignment method has been employed to align one target sequence to one template sequence **[19]**

(3) Computation of potentials due to mutation, gap penalty, secondary structure and solvent accessibility, pairwise interactions, triplet-interactions and quadruple interactions

Core part of the algorithm includes computing set of potentials that feed into the final score. The potentials were generated using different set of methods. The mutation potential values score for the amino acid sequence similarity between the probe sequence and the target fold. The values were computed by summing up the scores over the aligned, conserved secondary structural regions in the template. This gives a score for the substitution of an amino acid residue in the template sequence with that in the probe sequence. This score is obtained from the BLOSUM62 matrix **[42, 43].**

Mutation potential incorporated in HORIBALFRE was calculated using equations 1 and 2 **(see supplementary material).** The alignment between the sequence and structure will have gaps and a gap penalty was assigned after the alignment. An empirical gap opening penalty of 11 is chosen after examining the scores assigned to amino acid exchanges, while a gap extension is given a lesser penalty. The gaps in the secondary structure regions were penalized for more than a gap introduced in loop regions using equation 3 **(see supplementary material).**

Unlike comparative methods, which compare proteins based on sequence similarity and concept of homology, fold recognition methods take advantage of the extra information made available by 3D structure. A sum over the amino acid structural environment preferences over the entire sequence is a good indicator for the recognition of native-like folds **[36]**. The secondary structure details were mapped to the template sequence that is aligned to the probe sequence. The solvent accessibility values were mapped to the template sequence, according to the JOY-based structural feature definition **[44]**. The secondary structure and solvent accessibility values were paired to give a single score at each residue position. Eisenberg's 3D-1D substitution table **[39]** is then used to assign a score for the occupancy of 20 different amino acid residues at all the residue positions in the template. Hence, a score is obtained for the occupancy of an amino acid in the probe sequence into the corresponding, aligned residue position in the template. These values were summed up over the conserved and aligned regions between the sequence and the template. Hence, the combination of secondary structure and solvent accessibility values at different residue positions gives an environment score that is considered as a parameter. The results obtained only using environment scoring potential proves that subsets of sequence-structure pairs were not detected when the total potentials were considered. This explains the need for using weight factors for each of the parameters that were used for scoring. The potentials obtained for interactions between all residues. This interaction score thus depends on the observed frequency of interaction between two residues in already known protein structures. Pairwise potentials of mean force computed for a subset of protein structures derived from CAMPASS database is given in the additional material URL.

Pairwise, triplet and quadruple interactions were computed using HORI programs explained in our previous study **[41]**. The potentials were derived from a standard library of protein structures compiled from CAMPASS database **[15]**. The SCOP identifiers of the structures used to derive the potentials were given in supplementary material. Three-dimensional structure and amino acid sequence of proteins were related by an unknown set of rules that is often referred to as the folding code. This code is significantly influenced by non-local interactions between the residues **[34]**. If we approximate each residue as a sphere centered on its location, accordingly it is possible for three or even four closely packed spheres to make mutual contact, thus giving rise to three- or four-way interactions. Just as no more than four same-sized spheres can be in mutual contact in 3D space, no more than four–way interactions generally be expected to occur. We hypothesize that an interaction exists between two residues if the spatial distance between their C$^{\beta}$ atoms is within 7 Å and residues should be ≥4 residue positions apart in the template sequence. It is generally believed that the interactions involving loop residues can be ignored, as their contribution to fold recognition is relatively insignificant. In the current version of HORIBALFRE implementation, we consider only interactions between residues in the cores. It is observed that the scores obtained depend on the possible interactions within 7 Å and also all the interactions possible, depending on the residue pairs. For example, a positive potential was expected for Pro-Glu residue pair because the number of interactions possible within 7 Å for this pair is limited; only 0.8% of the total Pro-Glu interactions in the dataset. Similarly, as the number of possible Ala-Ala interactions is high, the potential is expected to be negative. In some sequence-structure pairs, no quadruple interaction potential value was obtained. But all protein structures showed pairwise interactions as expected, as the impact of constraints were less. Hence, inclusion of higher order interactions can make a distinction between sequence-structure pairs in the algorithm. Higher order residue interactions in HORIBALFRE algorithm were calculated using equations 4, 5, 6 and 7 **(see supplementary material)**.

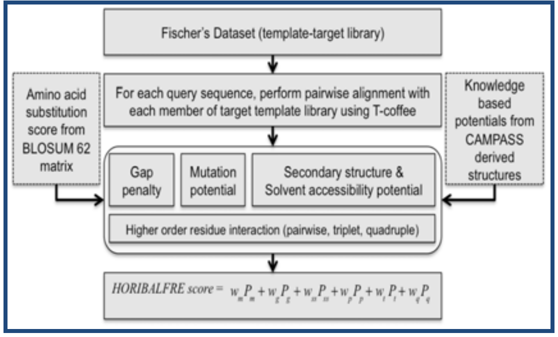

Fischer's Dataset (template-target library)
For each query sequence, perform pairwise alignment with each member of target template library using T-coffee
Amino acid substitution score from BLOSUM 62 matrix
Knowledge based potentials from CAMPASS derived structures
Gap penalty
Mutation potential
Secondary structure & Solvent accessibility potential
Higher order residue interaction (pairwise, triplet, quadruple)
HORIBALFRE score = $w_m P_m + w_g P_g + w_{ss} P_{ss} + w_p P_p + w_t P_t + w_q P_q$

Figure 1: Flowchart of HORIBALFRE algorithm. Algorithm derive parameters from multiple features like gap penalty, mutation potential, secondary structure and solvent accessibility potential, higher order residue interactions. Pre-computed potentials from CAMPASS dataset and BLOSUM 62 matrix were also incorporated.

Calculation of potentials of mean force using log-odds ratio:
Potentials of mean force were defined using pseudopotentials calculated from protein structures, pre-computed from a database using the inverse Boltzmann principle. Pairwise potentials of mean force have been used to study conformational ensembles **[11, 20]** .The potentials were obtained from the manually curated structures from the CAMPASS database. The formula used to calculate the log odds ratio is given in equations 8, 9 and 10 **(see supplementary material)**.

(4) Sum of Potentials (HORIBALFRE score)
The success of theoretical methods depends on the accuracy of the underlying scoring function that should be capable of

discriminating between correct (i.e. native) and incorrect configurations of the native polypeptide sequence **[4, 5].** HORIBALFRE score **(see equation 11 in supplementary material)** is derived using the summation of various potentials described earlier using the following equation. Additional weight factors were incorporated before the calculation of final HORIBALFRE score, and the values of the weight factors were determined based on the data available on the contribution of various potentials in protein structures **[38].**

(5) Ranking compatibility scores for sequence-fold pairs

The benchmarking was performed using the Fischer's dataset that consists of 68 sequence-structure pairs **[21].** The template-target pairs were analyzed to find out the considering only the mutation potential values. The mutation potential values obtained were comparatively lower in the cases where the right fold was identified. The mutation potential values showed dependence on the length of the sequences. In general, the potential values were high for sequences that differ hugely in length and with a poor sequence identity. As the difference in length increases, the negative values of mutation potentials were not observed. Mutation potential for a subset of template-target pairs is given in additional material URL.

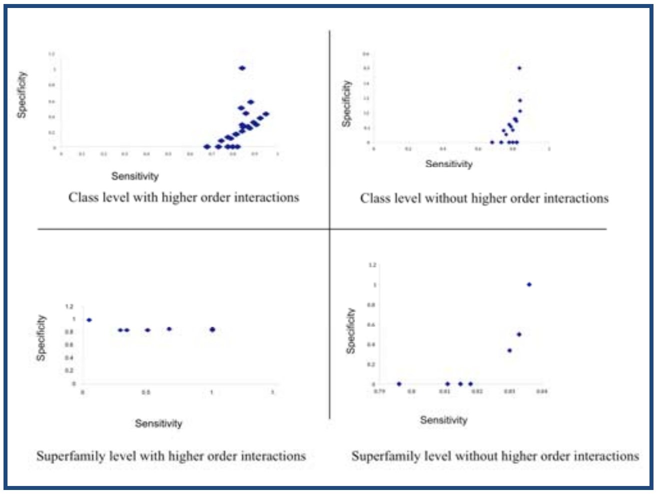
Statistical evaluation of HORIBALFRE algorithm:

Sensitivity and specificity analyses were performed at class level and at the superfamily level, with and without higher order residue interactions to illustrate the impact of higher order interactions in predicting the correct fold. Sensitivity and Specificity were calculated using equations 12 and 13 (**see supplementary material**).

Discussion:

HORIBALFRE score, an objective-scoring method introduced in this manuscript is calculated for 68 template-target members in the Fischer's dataset using a one-against-all method. Potentials explained in methods sections (See equations 1-10) were computed and used to derive HORIBALFRE score. Following the calculation of the HORIBALFRE score, a comparative analysis is performed using the scores without and with quadruple interactions. Representative template-target pairs without and with quadruple interactions are provided in **Table 1 and Table 2. (See supplementary materials)** Some of the sequence pairs were not observed amongst the top hits when quadruple interactions were not included. Here, we illustrate that higher order interactions contribute towards discriminating the correct fold amongst other folds, which give a similar score. The set of template-target pairs given in **Table 3,** (**See supplementary materials**) identified to have the corresponding fold pair amongst the top ten scores obtained for the sequence. From the class-wise distribution of the results, we observed that most of the folds that were predicted correctly (true positives) were belong to the mixed class of α/β. Further analysis will be required to elucidate whether quadruple interactions were biased towards specific SCOP classes. We had earlier shown that it is possible to discriminate between two folds of similar composition of supersecondary structures, the singly wound α/β barrels and doubly wound dehydrogenases, using higher order interactions **[41, also See additional Material URL].**

We used statistical validation that compared sensitivity and specificity analyses at the class level and superfamily level and analyzed the impact of the presence and absence of higher order residue interactions. For the sequence - pairs that were amongst the top 10 hits obtained using raw scores we calculated sensitivity and specificity. As expected, at the class level, the number of sequences that identified other structures that belonged to the same class as the native fold was higher than that obtained at the superfamily level. Class level with higher order interactions included, and superfamily level with higher order interactions are provided in **Figure 2**. As expected, at the class level, the number of sequences that identified other structures that belonged to the same class as the native fold was higher than that obtained at the superfamily level. The same data, when analyzed without the inclusion of quadruple and triplet interaction scores, it was seen that there were lesser number of sequences that identified other structures, belonging to the same class as the class of the native fold. The sensitivity at the class level also dropped to below 30%, while with the inclusion of higher order potentials, a higher sensitivity was obtained.



Class level with higher order interactions

Class level without higher order interactions

Superfamily level with higher order interactions

Superfamily level without higher order interactions

Figure 2: Sensitivity (*x-axis*) vs. Specificity (*y-axis*) plots based on HORIBALFRE results.

HORIBALFRE utilize three type of higher order residue interaction for characterizing correct fold for a given query sequence from a database of folds. We introduced various parameters incorporated in the algorithm and discussed pilot results in this manuscript. In an earlier study we showed that higher order residue interactions could delineate between closely related folds using two members from α/β folds from SCOP database **[41, 45, also See additional Material URL]**. The current results using 68 template-target pairs from Fisher's dataset used in HORIBALFRE indicates that fold recognition is being improved by the addition of higher order residue interaction potential due to quadruple interaction. The performance in the current analysis could be improved by inclusion of additional features. We will be applying normalization techniques and linear programming based function to improve the algorithm. Further, the algorithm will be tested on larger benchmark data sets to derive the coverage of algorithm and an integrated web server will be developed for fold prediction using higher order residue interactions.

Conclusion:

In this manuscript, we introduced and demonstrated results obtained from a novel fold recognition algorithm developed for protein fold recognition from sequence information. Fold recognition is an important problem in the current era of exponentially increasing sequenced genomes. Function and fold level annotation of newly sequenced genomes remains to be a priority. We designed a new fold recognition approach "HORIBALFRE" that utilize higher order residue interactions. The algorithm was tested using Fischer's dataset and performed a statistical evaluation. Preliminary results suggest that inclusion of higher order residue interactions, specifically quadruple interactions improves fold prediction.

Additional Material:
Additional datasets (PDB identifiers, Fischer's dataset, CAMPASS database derived structures), HORI-based pairwise, triplet and quadruple interaction scores and various parameters associated with HORIBALFRE scores and related programs are accessible from the URL: http://caps.ncbs.res.in/download/horibalfre/.

Acknowledgements:
S.P. was a Senior Research Fellow funded by the Ministry of Human Resource Development, Government of India and R.S. was a Senior Research Fellow funded by the Wellcome Trust, U.K. We would like to thank Department of Biotechnology, India for partial financial support. We also thank Indian Institute of Technology Roorkee and National Centre for Biological Sciences (TIFR) for infrastructural support.

References:

[1] Russ WP & Ranganathan R, *Curr Opin Struct Biol.* 2002 **12**: 447 [PMID: 12163066]
[2] Moult J & Melamud E, *Curr Opin Struct Biol.* 2000 **10**: 384 [PMID: 10851191]
[3] Jones DT et al. *Nature* 1992 **358**: 86 [PMID: 1614539]
[4] Anfinsen CB. *Biochem J.* 1972 **128**: 737 [PMID: 4565129]
[5] Anfinsen CB. *Science* 1973 **181**: 223[PMID: 4124164]
[6] Berman HM *et al. Nucleic Acids Res.* 2000 **28**: 235[PMID: 10592235]
[7] Lee D *et al. Nat Rev Mol Cell Biol.* 2007 **8**: 995[PMID: 18037900]
[8] Poole AM & Ranganathan R *Curr Opin Struct Biol.* 2006 **16**: 508[PMID: 16843652]
[9] David R *et al. Pharmacogenomics.* 2000 **1**: 445[PMID: 11257928]
[10] Jones DT *et al. Curr Opin Struct Biol.* 1996 **6**: 210[PMID: 8728653]
[11] Sippl MJ. *Curr Opin Struct Biol.* 1995 **5**: 229[PMID: 7648326]
[12] May AC *et al. Philos Trans R Soc Lond B Biol Sci.* 1994 **344**: 373[PMID: 7800707]
[13] Johnson MS *et al. Crit Rev Biochem Mol Biol.* 1994 **29**: 1[PMID: 8143488]
[14] Levinthal C J. *Chim Phys.* 1968 **65**: 44
[15] Sowdhamini R *et al. Structure.* 1998 **6**: 1094[PMID: 9753697]
[16] Notredame C *et al. J Mol Biol.* 2000 **302**: 205[PMID: 10964570]
[17] Chance MR *et al. Genome Res.* 2004 **14**: 2145[PMID: 15489337]
[18] Jones DT *et al. J Mol Biol.* 1999 **287**: 797[PMID: 10191147]
[19] Fischer D *et al. Pac Symp Biocomput.* 1996 : 300[PMID: 9390240]
[20] Sippl MJ *et al. J Mol Bio.* 1990 **213**: 859[PMID: 2359125]
[21] Go N. & Taketomi H. *Int J Pept Protein Res.* 1979 **13**: 235[PMID: 429100]
[22] Taketomi H *et al. Int J Pept Protein Res.* 1975 **7**: 445[PMID: 1201909]
[23] Liolios K *et al. Nucleic Acids Res.* 2006 **34**: 332[PMID: 17981842]
[24] Todd AE *et al. J Mol Biol.* 2001 **307**: 1113 [PMID: 11286560]
[25] Olszewski KA. *Pac Symp Biocomput.* 2000: 143[PMID: 10902164]
[26] Bhaduri A *et al. Proteins.* 2004 **54**: 657[PMID: 14997562]
[27] Duan MJ & Zhou YH. *Genomics Proteomics Bioinformatics.* 2005 3: 218[PMID: 16689689]
[28] Jiang N *et al. Int J Bioinform Res Appl.* 2006 **2**: 381[PMID: 18048179]
[29] Wyrwicz LS *et al. Acta Biochim Pol.* 2007 **54**: 551[PMID: 17882324]
[30] Dong E *et al. Gene.* 2008 **422**: 41[PMID: 18601985]
[31] Kokoszynska K *et al. Cell Cycle.* 2008 **7**: 2907[PMID: 18787404]
[32] Gu J *et al. J Comput Biol.* 2009 **16**: 427[PMID: 19254182]
[33] Xu Y *et al. J Comput Biol.* 1998 **5**: 597[PMID: 9773353]
[34] Tropsha A *et al. Pac Symp Biocomput.* 1996 : 614[PMID: 9390262]
[35] Bryant SH. *Proteins.* 1996 **26**: 172[PMID: 8916225]
[36] Bowie JU *et al. Science.* 1991 **253**: 164[PMID: 1853201]
[37] Bryant SH *et al. Proteins.* 1993 **16**: 92[PMID: 8497488]
[38] Godzik A et al. J Mol Biol. 1992 **227**: 227[PMID: 1522587]
[39] Wilmanns *M et al. Proc Natl Acad Sci U S A.* 1993 **90**: 1379[PMID: 8732766]
[40] Munson PJ *et al. Protein Sci.* 1997 **6**: 1467[PMID: 9232648]
[41] Sundaramurthy P *et al. BMC Bioinformatics.* 2010 **11**: S24. [PMID: 20122196]
[42] Altschul SF *et al. Nucleic Acids Res.* 1997 **25**: 3389[PMID: 9254694]
[43] Henikoff S & Henikoff JG, *Proc Natl Acad Sci U S A.* 1992 **89:** 10915[PMID: 1438297]
[44] Mizuguchi K *et al. Bioinformatics.* 1998 **14**: 617[PMID: 9730927]
[45] Sundaramurthy *P et al. Nature Protocols Network.* 2010. doi:10.1038/nprot.2010.91

Edited by P Kangueane
Citation: Sundaramurthy *et al.* Bioinformation 7(7): 352 - 359 (2011)

Supplementary material:

Equation	Equation number	Parameters
$E_m = \sum_{i=1}^{j} s[T(i),P(i)]$	(1)	Where *j* stands for the total number of positions in the aligned pair of sequences. *T(i)* stands for the target and *P(i)* is the probe. Position *i* should be within a regular secondary structural region in the template.
$E_s = \sum_{i=1}^{j} s[S(i),P(i)]$	(2)	Where *j* stands for the total number of positions in the aligned pair of sequences. *S(i)* gives the environment score for the position *i* and *P(i)* is the residue at position i in the probe sequence.
$E_g = \alpha + n \times \beta$	(3)	Where *a* is the gap opening penalty, *n* is the numbers of gaps and *β* is the gap extension penalty.
$E_{hori} = \sum E_{pairwise} + \sum E_{triplet} + \sum E_{quadruple}$	(4)	$E_{pairise}$ represents pseudoenergy due to pairwise interaction, $E_{triplet}$ represents pseudoenergy due to triplet interactions, $E_{quadruple}$ indicates pseudoenergy due to quadruple interactions. E_x (*x*, y) indicates pseudoenergy imparted by two interacting residues.
$E_{pairwise} = \sum_{i=1}^{j} (T(i),T(i+4))$	(5)	where *T(i)* is the i^{th} residue in the template sequence and *j* is the length of the sequence. The distance cut-off is 7 Å.
$E_{triplet}(i,j,k) = \sum E_p(i,j) + \sum E_p(j,k) + \sum E_p(k,i)$	(6)	Where *i, j, k* were ≥4 residues apart and the cut-off distance is 7 Å.
$E_{quadruple}(i,j,k,l) = \sum E_p(i,j) + \sum E_p(j,k) + \sum E_p(k,l) + \sum E_p(l,i) + \sum E_p(i,k) + \sum E_p(j,l)$	(7)	where *i,j,k,l* are each 4 residues apart and the cut-off distance is 7 Å.
$\Delta E(I,J) = -RT \ln[F(I,J) / P(I,J)]$	(8)	
$F(I,J) = f(i,j,d) / f(i,j)$	(9)	$f(i, j, d) =$ Occurrence of pair *i,j* within set distance d(here 7 Å) in the dataset; $f(i, j) =$ Total occurrences of pair *i,j* in the dataset;
$P(I,J) = \sum_{i}\sum_{j} p(i,j,d) / \sum_{i}\sum_{j}\sum_{d} p(i,j)$	(10)	$\sum\sum p(i,j,d) =$ Sum of over all possible pairwise interactions within 7 Å; $\sum\sum\sum p(i,j) =$ Sum over all possible pairwise interactions in the dataset without any distance cut-off. The temperature was set to 293K, so that RT = 0.582 *kcal/mol*.
HORIBALFRE score = $w_m P_m + w_g P_g + w_{ss} P_{ss} + w_p P_p + w_t P_t + w_q P_q$	(11)	*Where,* P_m is the sum of potentials due to mutation; P_g is sum of potentials due to gap penalty; P_{ss} is sum of potentials due secondary structure compatibility and solvent accessibility; P_p is sum of potentials due to pairwise interactions; P_t is sum of potentials due to triplet interactions; P_q is sum of potentials due to quadruple interactions; and w_m = 0.1; w_g =0.1; w_{ss} = 0.2; w_p =0.1; w_t=0.2;w_q = 0.3
$\text{Sensitivity} = \frac{TP}{TP + FN}$	(12)	Where, TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.
$\text{Specificity} = \frac{TN}{TN + FP}$	(13)	

Table 1: HORIBALFRE scores obtained without inclusion of quadruple interactions.

Sequence-Fold pairs	Template length	Target length	HORIBALFRE Score
1chra_2mnr	370	357	113.409
1isua_2hipa	62	72	48.286
8i1b_4fgf.	152	146	57.152
1mup_1rbp	166	182	83.570
1fxia_1ubq	96	76	61.686
3chy_4fxn	128	138	29.921
2azaa_1paz	129	123	3.040
1sim_1nsba	381	390	104.098

Table 2: HORIBALFRE scores obtained with inclusion of quadruple interactions.

Sequence-Fold pairs	Template length	Target length	HORIBALFRE Score
1npx_3grs	447	478	264.973
1bbha_2ccya	131	128	87.648
1chra_2mnr	370	357	113.409
1isua_2hipa	62	72	48.430
8i1b_4fgf	152	146	58.185
1mup_1rbp	166	182	83.570
1stfb_1mola	98	87	74.551
1fxia_1ubq	96	76	63.600
3chy_4fxn	128	138	29.921
2azaa_1paz	129	123	8.979
1sim_1nsba	381	390	105.836

Table 3: Set of template-target pair identified by HORIBALFRE with corresponding folds pair amongst the top ten scores obtained for the sequence.

Sequence and Fold	Potential	Class ID	Fold ID	Superfamily ID	Family ID	Class	Fold	Superfamily
1npx_3grs	239.816	30441	51349	51904	51905	Alpha and beta proteins (a/b)	FAD/NAD(P)-binding domain	FAD/NAD(P)-binding domain
1bbha_2ccya	73.222	16544	46456	47161	47175	All alpha proteins	Four-helical up-and-down bundle	Cytochromes
1chra_2mnr	28.220	29245	51349	51350	51604	Alpha and beta proteins (a/b)	TIM beta/alpha-barrel	Enolase C-terminal domain-like
1gky_3adk	96.625	31885	51349	52539	52540	Alpha and beta proteins (a/b)	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
2hhma_1fbpa	174.581	42916	56572	56654	56655	Multi-domain proteins (alpha and beta)	Carbohydrate phosphatase	Carbohydrate phosphatase
1isua_2hipa	43.695	44991	56992	57651	57652	Small proteins	HIPIP (high potential iron protein)	HIPIP (high potential iron protein)
1gal_3cox	460.067	30332	51349	51904	51905	Alpha and beta proteins (a/b)	FAD/NAD(P)-binding domain	FAD/NAD(P)-binding domain
1mioc_1minb	360.606	35610	51349	53799	53807	Alpha and beta proteins (a/b)	Chelatase-like	"Helical backbone" metal receptor
8i1b_4fgf	55.457	25486	48724	50352	50353	All beta proteins	beta-Trefoil	Cytokine
1mup_1rbp	79.343	27085	48724	50813	50814	All beta proteins	Lipocalins	Lipocalins
1cpcl_1cola	61.927	43378	56835	56836	56837	Membrane and Cell	Toxins' membrane translocation	Colicin

						surface proteins and peptides	domains	
1atna_1atr	162.765	33421	51349	53066	53067	Alpha and beta proteins (a/b)	Ribonuclease H-like motif	Actin-like ATPase domain
1arb_4ptp	119.488	25831	48724	50493	50494	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases
1ltsd_1bova	72.833	25070	48724	50198	50203	All beta proteins	OB-fold	Bacterial enterotoxins
1stfi_1mola	76.258	37988	53931	54402	54403	Alpha and beta proteins (a+b)	Cystatin-like	Paramphistomum epiclitum [TaxId: 54403]
1fxia_1ubq	50.160	37585	53931	54235	54236	Alpha and beta proteins (a+b)	beta-Grasp (ubiquitin-like)	Ubiquitin-like
3hlab_2rhe	89.066	20523	48724	48725	48726	All beta proteins	Immunoglobulin-like beta-sandwich	Immunoglobulin
3chy_4fxn	17.739	31197	51349	52171	52218	Alpha and beta proteins (a/b)	Flavodoxin-like	Flavoproteins
2azaa_1paz	0.656	22878	48724	49502	49503	All beta proteins	Cupredoxin-like	Cupredoxins
1cew_1mola	83.485	37988	53931	54402	54403	Alpha and beta proteins (a+b)	Cystatin-like	Paramphistomum epiclitum [TaxId: 54403]
1sim_1nsba	105.284	27597	48724	50938	50939	All beta proteins	6-bladed beta-propeller	Sialidases
1gp1a_2trxa	91.403	32719	51349	52832	52833	Alpha and beta proteins (a/b)	Thioredoxin fold	Thioredoxin-like