BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Incorporating background frequency improves entropy-based residue conservation measures

Kai Wang and Ram Samudrala*

Address: Computational Genomics Group, Department of Microbiology, University of Washington, USA

Email: Kai Wang - dna@u.washington.edu; Ram Samudrala* - ram@compbio.washington.edu

* Corresponding author

Published: 17 August 2006

BMC Bioinformatics 2006, **7**:385 doi:10.1186/1471-2105-7-385

Received: 14 June 2006

Accepted: 17 August 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/385

Abstract

Background: Several entropy-based methods have been developed for scoring sequence conservation in protein multiple sequence alignments. High scoring amino acid positions may correlate with structurally or functionally important residues. However, amino acid background frequencies are usually not taken into account in these entropy-based scoring schemes.

Results: We demonstrate that using a relative entropy measure that incorporates amino acid background frequency results in improved performance in identifying functional sites from protein multiple sequence alignments.

Conclusion: Our results suggest that the application of appropriate background frequency information may lead to more biologically relevant results in many areas of bioinformatics.

Background

Protein multiple sequence alignments are widely used to infer conservation of amino acid residues within an evolutionarily related family [1,2]. Highly conserved residues tend to correlate with structural and/or functional importance, and accurate identification of such important residues aids in experimental characterization of protein function.

One commonly used sequence conservation measure is the entropy score. In its simplest form, the entropy score for each aligned column in a multiple sequence alignments can be expressed as:

$$S_{entropy} = H = -\sum_{i}^{n_{aa}} p_i \log_2 p_i \qquad (1)$$

where n_{aa} is the number of residue types in the column representing an alignment position, and p_i represents the observed frequency of residue type i in the aligned column. Other, more complicated, residue conservation measures derived from the entropy score have been developed and used in identifying functionally important residues [3-8].

In recent years many sophisticated functional site prediction algorithms have been developed (for reviews, see [9,10]). Many of these prediction algorithms implicitly or explicitly analyze the amino acid variations in a given position in multiple alignments. The evolutionary trace method analyzes residue variation patterns within and between protein subfamilies from multiple alignments, and maps important residues to protein structure [11,12]. A further development of this method incorporates entropy information for more accurate ranking of residue importance [13]. Oliveira *et al* devised entropy-variability

plots that can be used to identify structurally and functionally important residues [14]. Pei *et al* used the conservation difference between artificial sequence profiles and naturally occurring sequence profiles to detect homology and identify active sites [15]. Soyer *et al* used site-specific evolutionary models for predicting functional sites in proteins [16]. Wang *et al* used linear models to analyze multiple alignments for mutated viral proteins to identify amino acid positions important for drug resistance [17]. Chelliah *et al* identified interaction sites by separating the structural and functional constraints for each position in multiple alignments [18]. Greaves *et al* used geometry-based and sequence profile-based calculation for predicting enzyme active sites [19]. Cheng *et al* introduced a hybrid method incorporating both sequence conservation and structural stability to predict functional sites [20]. The SIFT server automatically constructs multiple alignments for query sequence and then predicts amino acid substitutions that are likely to affect protein function [21]. The ConSurf server identifies protein functional region by surface mapping of phylogenetic information inferred from multiple alignments [22]. The MINER server uses phylogenetic motifs from multiple alignments to identify protein functional site regions [23]. Besides functional site identification, the residue conservation information can also be used to identify residues determining subfamiliy functional specificity [24-28]. The prevalence of these methods suggests that the extraction of conservation information from multiple alignments is important for the correct prediction of functional residues or regions.

Results and discussion

Rationale

Here, we argue that entropy scores that do not incorporate background amino acid frequencies are not theoretically optimal for calculating residue conservation. To demonstrate this, we rewrite the entropy score using a uniform amino acid frequency distribution P_u ($p_u = 1/n_{aa}$ for each residue type in the aligned column):

$$\begin{aligned} H &= -\sum_i^{n_{aa}} p_i \log_2 p_i \\ &= -\sum_i^{n_{aa}} p_i (\log_2 p_i - \log_2 p_u) - \sum_i^{n_{aa}} p_i \log_2 p_u \\ &= -\sum_i^{n_{aa}} p_i \log_2 (p_i / p_u) - \log_2 p_u \\ &= -D(P_i \,||\, P_u) - \log_2 p_u \end{aligned} \qquad (2)$$

where $D(P_i \,||\, P_u)$ is commonly referred to as the relative entropy or Kullback-Leibler divergence. Therefore, the entropy score is numerically identical to the negative relative entropy between the observed amino acid frequency distribution P_i and a uniform distribution P_u, minus a constant. In statistics, relative entropy arises as the expected logarithm of the likelihood ratio, and it may be used as a measure of the distance between two probability distributions [29]. In the case of residue conservation, the higher deviation from the "background" indicates stronger evolutionary constraint, which suggests that this position may perform an important functional role. However, nature does not sample every amino acid equally when creating proteins. Therefore, the simple uniform distribution P_u above is not optimal as a reference distribution to evaluate functional importance.

We propose that a relative entropy measure incorporating the observed background frequency from protein sequence databases would be a better measure to capture the functional importance of amino acid residues. More specifically, we propose to use the formula:

$$S_{relative\_entropy} = D(P_i \,||\, P_{ib}) = \sum_i p_i \log_2 (p_i / p_{ib}) \qquad (3)$$

where P_{ib} can represent the background amino acid frequencies found in naturally occurring protein sequences, or any other arbitrary set of background frequencies. This measure will increase the scores for aligned columns containing "rare" residues, which are often functionally important. To explain this in a more intuitive way, consider two invariant positions that have only cysteines and only serines, respectively. The position with cysteines is more likely to be functional. The entropy measure will assign the same score to the two positions, but the relative entropy measure assigns a higher score to the invariant cysteine position, since cysteine has a much lower background frequency (~2%) than serine (~7%).

Comparative analysis of entropy score and relative entropy score

To investigate whether our proposed relative entropy measure ($S_{relative\_entropy}$) is more sensitive than the entropy measure ($S_{entropy}$) in detecting functional sites, we evaluated the performance of these measures to identify functionally important residues using the Thornton [30] and Lovell datasets [18]. In addition, to investigate how the use of different background frequencies affects the performance of the relative entropy method, we used two sets of frequencies: the general background frequencies observed in nature and the family-specific background frequencies retrieved from the alignments for each query sequence.

For each protein in the datasets, we built multiple alignments and calculated the entropy or relative entropy scores for each residue and evaluated their performance by two criteria: The first criterion is the ROC score, which measures how the quantitative scores correlate with true

functional sites. The second criterion is the "top 10 hits" score, which counts the number of functionally important residues that are in the top 10 highest scoring residues. We found that the relative entropy method significantly outperforms the entropy method under both evaluation criteria for both datasets (Figure 1). In addition, using family-specific background frequencies in the relative entropy method has similar performance to using general background frequencies.

To further investigate why the relative entropy methods perform better than the entropy method, we selected an example from the Thornton dataset where both the relative entropy methods accurately identify the active sites in their "top 10 hits", but where the entropy method fails to do so. This example protein (PDB identifier 1a65A) is an oxidoreductase, and contains three consecutive active sites (His-Cys-His). We plotted the structures as ribbon representation and colored the residues by their scores

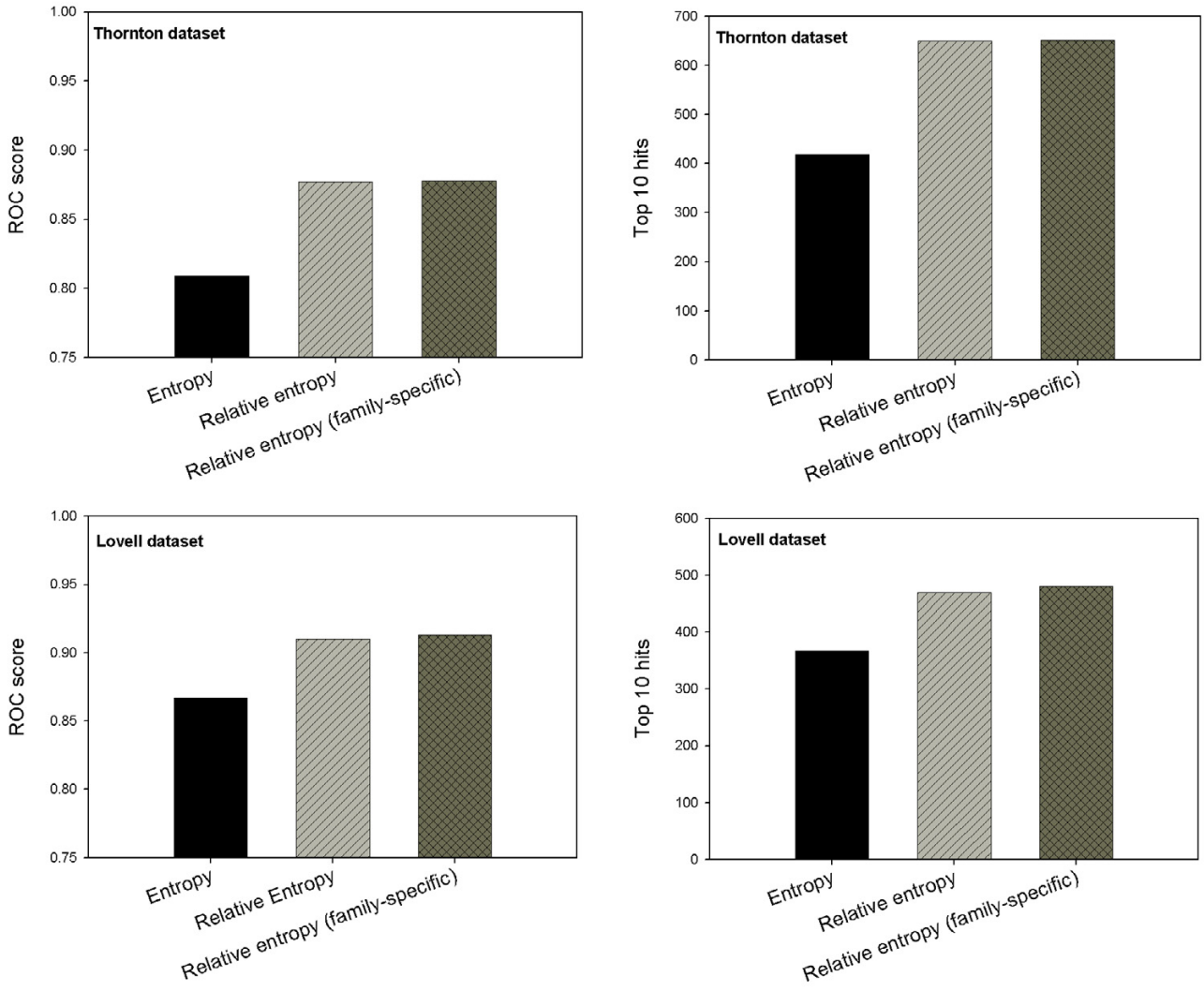

Figure 1
Accuracy of functional site identification by the entropy method, the default relative entropy method and the family-specific relative entropy method on the Thornton and Lovell datasets. The average ROC score and the top 10 hits score are used to evaluate performance. The family-specific relative entropy method has negligible difference in performance compared to the default relative entropy method using general background frequencies. Both relative entropy methods perform better than the entropy method, demonstrating that the incorporation of background frequency information improves functional site identification.

from each of the three methods (red color indicating high scoring positions) (Figure 2). We found that the entropy method incorrectly assigns the highest scores to the region near the C-terminal of the protein (Asp-Asp-Leu-Pro-Pro-Glu-Ala-Thr-Ser-Ile-Gln-Thr-Val) and not for the residues in the active site (His-Cys-His, depicted as spheres in figure). The frequencies of these amino acids in nature are approximately 5%, 9%, 5%, 6%, 8%, 6%, 7%, 5%, 4% and 6% for Asp, Leu, Pro, Glu, Ala, Thr, Ser, Ile, Gln and Val, respectively, but only 2% and 2% for His and Cys, respectively. Therefore, by applying the relative entropy measure, we down-weighted the C-terminal region and predicted higher scores for the active sites. For the family-specific relative entropy method, the family-specific frequencies are 4% for His and 1% for Cys, so the position with the Cys has a higher score (more intense red color) than the neighboring His, but all three sites are still among the top 10 hits. Our analysis indicates that taking into account background frequencies boosts the scores for positions containing rare amino acids and results in improved performance for identification of functionally important positions.

Comparative analysis of entropy scores using more accurate frequency estimates

We further investigated whether entropy-based methods could benefit from incorporating more accurate amino acid frequency information. We compiled a HMM model from multiple alignments for each query sequence, and calculated the positional entropy or relative entropy (using the general background frequencies) for each aligned column, similar to a previous study [24]. There are two main advantages of using HMM models: (1) sequences are weighted so that the effects of uneven or biased database sampling are reduced and (2) frequencies of unobserved amino acids can be estimated through the use of Dirichlet mixtures [31].

We evaluated the performance of the HMM-derived entropy and relative entropy methods, as well as several other residue conservation measures, including three AL2CO-based methods [2] and the SCORECONS method [32]. AL2CO is a program that implements an entropy-based method (AL2CO_entropy), a variance-based method (AL2CO_variance) and a sum-of-pairs method (AL2CO_sop). The AL2CO_entropy method is similar to our entropy method; the AL2CO_variance method uses background frequencies that are estimated from the alignment; and the AL2CO_sop method uses pairwise similarity scores derived from a given amino acid substitution matrix. All three AL2CO-based methods apply the default independent count weighting scheme [33] to weight sequences in alignments. The SCORECONS method generates a composite score that takes into account amino acid frequencies, stereochemical diversity, gap penalties, and sequence weighting. This scoring method has been

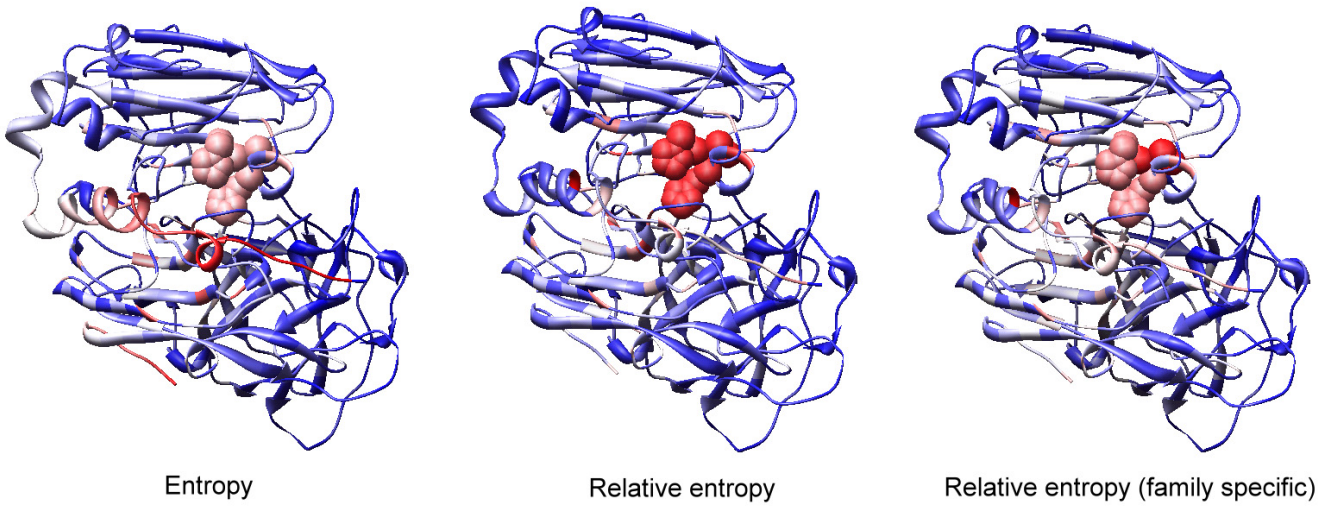

Figure 2
Example comparison of relative entropy and entropy methods. The structure of the example protein, an oxidoreductase (PDB identifier 1a65A), is shown in ribbon representation with the functional active sites (His-Cys-His) represented as spheres. For each method evaluated, each residue in the structure is colored by its predicted functional importance score, with the color changing from red to white to blue as the score decreases. The entropy method incorrectly assigns the highest scores to residues in the C-terminal region, but the two relative entropy methods correctly assign the highest scores to the active sites.

used as a benchmarking score for protein-DNA interaction site identification [32] and protein functional site identification [20]. We found that when HMM-derived amino acid frequencies are used, the relative entropy method still outperforms the entropy method (Figure 3), and both methods outperform the three AL2CO-based methods and the SCORECONS method. Among the three AL2CO-based methods, the AL2CO_variance method performs the best, which may be partially due to its use of background frequencies. The SCORECONS method, which does not outperform the HMM-derived entropy method, was originally developed for scoring conservation in protein-DNA interaction sites and therefore may not be well-suited for predicting functionally important residues in general. In summary, our analysis suggests that using more accurate amino acid frequency estimates, together with using appropriate background frequencies, results in improved functional site prediction from multiple alignments.

Improvements for functional site prediction can occur by increasing the sophistication of the measures considered: The relative entropy method only scores individual positions without considering neighboring residues. A method that analyzes context (neighboring residues) may improve performance. The multiple alignments generated

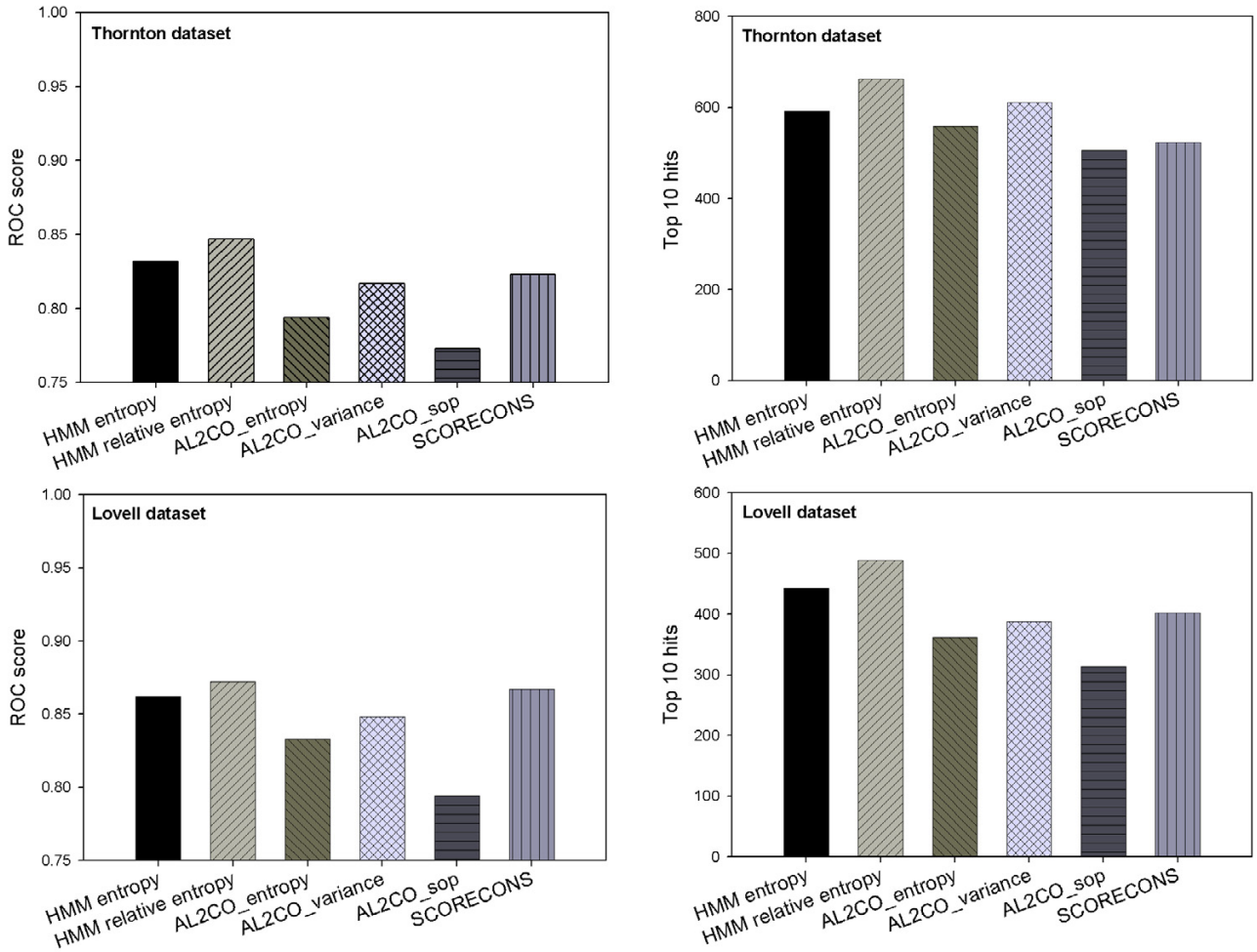

Figure 3
Accuracy of functional site identification by the HMM-derived entropy method, the HMM-derived relative entropy method, three AL2CO-based methods (AL2CO_entropy, AL2CO_variance and AL2CO_sop) and the SCORECONS method on the Thornton and Lovell datasets. The average ROC score and the top 10 hits score are used to evaluate performance. The HMM-derived relative entropy method has the best performance, demonstrating the importance of using background frequency information, as well as accurate estimates of amino acid frequencies.

by the PSI-BLAST program may not be optimal, and more accurate multiple alignments (such as those generated by the HMM method) may improve performance. In addition, an appropriate treatment of gaps and sequence weights (for example, a family specific treatment), consideration of phylogenetic relationships among sequences, and analysis of local structural information when available, is also likely to improve performance. We believe that a hybrid method that incorporates the relative entropy method and all the above improvements will have significantly better performance for functional site prediction.

Conclusion

In conclusion, the use of background frequency information significantly improves entropy-based functional site prediction. This principle has been advocated before (such as in [34]), but its use has been very limited: For example, sequence logo is widely used to visually display conserved nucleotide or amino acid sites in sequence motifs; however, many logo generation programs [35-38] are unable to accept user-supplied background frequencies. (Some exceptions include the PICTOGRAM server [39] and the CONSENSUS server [40].) In addition, several programs for editing or annotating multiple alignments exist [41,42], but they are unable to use background frequencies to calculate relative entropy for each aligned residue.

The use of background information is limited in other areas of bioinformatics as well. For example, many programs are available to identify "functional enrichment" for a list of genes from microarray experiment, but only a few of them are able to accept a user-supplied list of "background genes". Dozens of tools are available to identify transcription factor binding sites (TFBS) or other functional motifs in a given sequence, but few of them are able to take into account the background frequency of predicted TFBS or motifs in the corresponding genome (exceptions include [43]). Fold recognition methods are widely used to assign a query sequence to a structural fold, but few considers the relative abundance (or prior probability) of candidate folds in the corresponding proteome. The broader application of appropriate background information in all areas of bioinformatics will lead to more biologically relevant results.

Methods

Data sources

We used two datasets consisting of protein functional sites for evaluating different algorithms. The Thornton dataset was compiled manually from primary literature on known protein structures and was shown to be more comprehensive and specific than the SITE annotation in PDB files [30]. The Lovell dataset contains manually compiled protein functional sites, including ligand binding sites and enzyme active sites [18]. The Thornton dataset contains 1,546 enzyme active sites from 508 proteins, and the Lovell dataset contains 1,137 functional sites from 243 proteins. The same versions of the two sets have been used previously for structure-based searching of functional sites [20].

Criteria for performance evaluation

We used two criteria for evaluating the performance of functional site prediction algorithms. The first criterion is the ROC score, which computes the area under a curve that plots fraction of true positives versus false positives by varying the threshold value of classification. A perfect classification algorithm that puts all the functional sites at the top of the ranked residue list has an ROC score of 1, and a random classification algorithm has an ROC score of 0.5. The second criterion is the "top 10 hits" score, which computes the number of functional sites among the top 10 scoring residues in a given protein. A perfect classification algorithm has a top 10 hits score of N_{fun} or 10 (if N_{fun} is more than 10), while a random classification algorithm has an average top 10 hits score of $10*N_{fun}/N$ ($N_{fun}$and N denote the number of functional residues and the number of all residues in the given protein, respectively).

Functional site identification

We evaluated several functional site identification methods that take protein multiple sequence alignments as input for their predictions. For the Thornton and Lovell datasets, we generated multiple sequence alignments by searching each query sequence against the Uniref90 database [44] with the PSI-BLAST program blastpgp in the BLAST program package [45]. We used three iterations (through the "-j 3" option), the "-m 6" display option and all other default parameters for the PSI-BLAST program. The resulting multiple sequence alignments were converted to ClustalW format, and then analyzed by various scoring methods described below.

For the entropy and relative entropy method, we used equation (1) and (3) in the main text to assign a score to each aligned column in the multiple alignments. We treated gaps in the same way as an amino acid, though we found little performance difference when ignoring gaps. Only the residue types that appear in aligned columns were used in the computation of the relative entropy score. The background distribution P_{ib} in equation (3) can be varied to explore the use of different background frequencies. In our benchmarking experiment, we used two different sets of background frequencies: (1) the general background frequencies defined in karlin.c program of the BLAST package [45], and (2) the family-specific background frequencies observed in the multiple alignments for each query sequence.

For the HMM-derived entropy and relative entropy method, we first compiled a HMM model using the hmmbuild program in the HMMER package [46] with default parameters, and then calculated the positional entropy or relative entropy using amino acid frequencies estimated by the HMM model. The general background frequencies are used for the relative entropy computation.

We also used three conservation measures implemented in the AL2CO program [2], including the entropy measure (AL2CO_entropy), the variance measure (AL2CO_variance) and the sum-of-pairs measure (AL2CO_sop). All the three methods use the default "independent count" scheme for weighting sequences [33]. The default parameters for the AL2CO program were used for computation, except that the BLOSUM62 matrix rather than identity matrix was used in the sum-of-pairs method. For the SCORECONS method [32], we used the scorecons program with default parameters.

Authors' contributions

KW carried out the computational experiments and drafted the manuscript. RS developed the idea, provided intellectual guidance and mentorship. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by a Searle Scholar Award, a NSF CAREER award, NSF grant DBI-0217241, and NIH grant GM068152-01. We wish to thank members of the Samudrala group and Dr. Sridhar Hannenhalli for helpful discussions and comments.

References

1. Valdar WS: **Scoring residue conservation.** *Proteins* 2002, **48(2):**227-241.
2. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment.** *Bioinformatics* 2001, **17(8):**700-712.
3. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9(1):**56-68.
4. Shenkin PS, Erman B, Mastrandrea LD: **Information-theoretical entropy as a measure of sequence variability.** *Proteins* 1991, **11(4):**297-313.
5. Gerstein M, Altman RB: **Average core structures and variability measures for protein families: application to the immunoglobulins.** *J Mol Biol* 1995, **251(1):**161-175.
6. Williamson RM: **Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters.** *J Theor Biol* 1995, **174(2):**179-188.
7. Mirny LA, Shakhnovich EI: **Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function.** *J Mol Biol* 1999, **291(1):**177-196.
8. Plaxco KW, Larson S, Ruczinski I, Riddle DS, Thayer EC, Buchwitz B, Davidson AR, Baker D: **Evolutionary conservation in protein folding kinetics.** *J Mol Biol* 2000, **298(2):**303-312.
9. Jones S, Thornton JM: **Searching for functional sites in protein structures.** *Curr Opin Chem Biol* 2004, **8(1):**3-7.
10. Watson JD, Laskowski RA, Thornton JM: **Predicting protein function from sequence and structural data.** *Curr Opin Struct Biol* 2005, **15(3):**275-284.
11. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257(2):**342-358.
12. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O: **An accurate, sensitive, and scalable method to identify functional sites in protein structures.** *J Mol Biol* 2003, **326(1):**255-261.
13. Mihalek I, Res I, Lichtarge O: **A family of evolution-entropy hybrid methods for ranking protein residues by importance.** *J Mol Biol* 2004, **336(5):**1265-1282.
14. Oliveira L, Paiva PB, Paiva AC, Vriend G: **Identification of functionally conserved residues with the use of entropy-variability plots.** *Proteins* 2003, **52(4):**544-552.
15. Pei J, Dokholyan NV, Shakhnovich EI, Grishin NV: **Using protein design for homology detection and active site searches.** *Proc Natl Acad Sci U S A* 2003, **100(20):**11361-11366.
16. Soyer OS, Goldstein RA: **Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters.** *J Mol Biol* 2004, **339(1):**227-242.
17. Wang K, Jenwitheesuk E, Samudrala R, Mittler JE: **Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance.** *Antivir Ther* 2004, **9(3):**343-352.
18. Chelliah V, Chen L, Blundell TL, Lovell SC: **Distinguishing structural and functional restraints in evolution in order to identify interaction sites.** *J Mol Biol* 2004, **342(5):**1487-1504.
19. Greaves R, Warwicker J: **Active site identification through geometry-based and sequence profile-based calculations: burial of catalytic clefts.** *J Mol Biol* 2005, **349(3):**547-557.
20. Cheng G, Qian B, Samudrala R, Baker D: **Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design.** *Nucleic Acids Res* 2005, **33(18):**5861-5867.
21. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31(13):**3812-3814.
22. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N: **ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W299-302.
23. La D, Livesay DR: **Predicting functional sites with an automated algorithm suitable for heterogeneous datasets.** *BMC Bioinformatics* 2005, **6:**116.
24. Hannenhalli SS, Russell RB: **Analysis and prediction of functional sub-types from protein sequence alignments.** *J Mol Biol* 2000, **303(1):**61-76.
25. Vilim RB, Cunningham RM, Lu B, Kheradpour P, Stevens FJ: **Fold-specific substitution matrices for protein classification.** *Bioinformatics* 2004, **20(6):**847-853.
26. Bielawski JP, Yang Z: **A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution.** *J Mol Evol* 2004, **59(1):**121-132.
27. Pei J, Cai W, Kinch LN, Grishin NV: **Prediction of functional specificity determinants from protein sequences using log-likelihood ratios.** *Bioinformatics* 2006, **22(2):**164-171.
28. Mirny LA, Gelfand MS: **Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors.** *J Mol Biol* 2002, **321(1):**7-20.
29. Cover TM, Thomas JA: **Elements of information theory.** In *Wiley series in telecommunications* Edited by: Schilling DL. New York , John Wiley & Sons; 1991.
30. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32 Database issue:**D129-33.
31. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D: **Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology.** *Comput Appl Biosci* 1996, **12(4):**327-345.
32. Valdar WS, Thornton JM: **Protein-protein interfaces: analysis of amino acid conservation in homodimers.** *Proteins* 2001, **42(1):**108-124.
33. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN: **PSIC: profile extraction from sequence alignments with position-specific counts of independent observations.** *Protein Eng* 1999, **12(5):**387-394.
34. Stormo GD: **Information content and free energy in DNA--protein interactions.** *J Theor Biol* 1998, **195(1):**135-137.
35. Schuster-Bockler B, Schultz J, Rahmann S: **HMM Logos for visualization of protein families.** *BMC Bioinformatics* 2004, **5:**7.

36. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14(6):**1188-1190.
37. Bindewald E, Schneider TD, Shapiro BA: **CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments.** *Nucleic Acids Res* 2006, **34(Web Server issue):**W405-11.
38. Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV: **enoLOGOS: a versatile web tool for energy normalized sequence logos.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W389-92.
39. PICTOGRAM: **[http://genes.mit.edu/pictogram.html].** .
40. CONSENSUS: **[http://adric.wustl.edu/oldconsensus].** .
41. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor.** *Bioinformatics* 2004, **20(3):**426-427.
42. Johnson JM, Mason K, Moallemi C, Xi H, Somaroo S, Huang ES: **Protein family annotation in a multiple alignment viewer.** *Bioinformatics* 2003, **19(4):**544-545.
43. Levy S, Hannenhalli S: **Identification of transcription factor binding sites in the human genome sequence.** *Mamm Genome* 2002, **13(9):**510-514.
44. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucleic Acids Res* 2006, **34(Database issue):**D187-91.
45. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
46. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14(9):**755-763.