Information content of sets of biological sequences revisited

Alessandra Carbone and Stefan Engelen

Génomique Analytique, Université Pierre et Marie Curie, INSERM UMRS511, 91, Bd de l'Hôpital, 75013 Paris, France; e-mail: Alessandra.Carbone@lip6.fr and stefengelen@gmail.com

Abstract

To analyze the information included in a pool of amino-acid sequences, a first approach is to align the sequences, to estimate the probability of each amino-acid to occur within columns of the aligned sequences and to combine these values through an "entropy" function whose minimum corresponds to absence of information, that is to the case where each amino-acid has the same probability to occur. Another alternative is to construct a distance tree between sequences (issued by the alignment) based on sequence similarity and to properly interpret the tree topology so to model the evolutionary property of residue conservation. We introduced the concept of "evolutionary content" of a tree of sequences, and demonstrated at what extent the more classical notion of "information content" on sequences approximates the new measure and in what manner tree topology contributes sharper information for the detection of protein binding sites.

INTRODUCTION

Comparison of multiple amino-acid sequences resulting from years of evolution demonstrated to provide insightful information on the relationships between sequence, structure, function and evolution of protein families (Lecompte et al. 2001; Duret and Abdeddaim, 2000). Multiple sequence alignments were originally used to explore phylogenetic relationships between organisms (Phillips et al. 2000), and more recently, to detect more and more distant homologues, conserved functional features of proteins and major evolutionary events in genomes (Altschul et al. 1997; Thompson et al. 1999; Wallace et al. 2005; Notredame 2002; Notredame 2007). Also, significant improvements in predictions of both 3D fold (Moult 2005) and function (Watson et al. 2005) are also achieved through multiple sequence comparison.

Pools of aligned amino-acid sequences are usually constituted by very few sequence instances which are available around us today, and their grouping in sequence space highlights potential similar sequences which might exist but that we did not "see" (yet). We argue that a definition of the information content of a tree issued from a sequence alignment has to be based on the information coming from these potential viable sequences too. In this study we check the hypothesis that the topology of a distance tree of sequences codes for interesting "biological" information of evolutionary origin, which can be extracted from a combinatorial analysis of the tree and a suitable interpretation of its nodes. To establish the conditions under which a pool of sequences organized in a tree is informative or not is a primary question addressed by our model.

The information content of a biological molecule corresponds to the number of sequence constraints that have to be conserved to maintain its function under random mutations (Adami and Cerf 2000; Carothers et al. 2004). Expressed in amino-acid units, the maximum information content that can be encoded on a N-amino-acid long protein sequence is precisely N, which defines a unique sequence among the 20^N different protein sequences with N amino-acids. The calculation of the maximum information content encoded on a tree of N leaves, is precisely N when we admit the leaves to be labeled by the same sequence. If the more realistic hypothesis of considering leaves labeled by different sequences is adopted, then the computation is more subtle since it depends on the topology of the tree. In this paper we explain how to make this computation.

In what follows, sequences are made out of amino-acids a_k, with $k = 1 \dots 20$. The approach to multiple sequence analysis that we propose is general and it might be applied to arbitrary finite languages. An important general insight to retain from this analysis is that what we "observe" (that is, the actual data) is just a small amount of what we actually represent through trees constructed with clustering algorithms that group together biological objects by similarity.

THE MODEL

We suppose to have N sequences which have been aligned, L be the length of the alignment and T be the associated distance tree, whose N leaves are labeled by sequences. We think of the root of T as the set of all possible sequences "represented" by the N sequences of the original pool as follows. For each position i in the alignment, we define a characteristic function

$$\chi_i(a_k) = \begin{cases} 0 & \text{no residue } a_k \text{ appears at position } i \\ 1 & \text{otherwise} \end{cases}$$

where $k = 1 \dots 20$. Since an alignment contains gaps, we encode when needed, a gap as a 21st residue and we name it a_{21}. We let $P'_0 = \prod_{i=1}^{L} \sum_{k=1}^{21} \chi_i(a_k)$ be the number of potential sequences which are *coherent* with the original pool of sequences, that is those sequences which are composed by residues which appear at least once at a given position. Note that having considered a gap as a residue, we count here also aligned sequences which are formed by gaps at almost all positions. These sequences might be considered undesired and if so, it is reasonable to substract from the pool P'_0 all sequences containing more than $\frac{L}{2} - 1$ gaps. This way, sequences cover at least the 50% of a sequence alignment and for any two sequences in the alignment the overlap is guaranteed. We call P_0 the cardinality of the resulting set of potential sequences.

As done for the root, a value P can be associated to any internal node of the tree. It corresponds to all potential sequences represented by the sequences labeling the leaves of the associated subtree.

We are interested to evaluate the information content of the pool of aligned sequences at a position i which is induced by the tree structure, where $i = 1 \dots L$. For each i, we consider the S^i maximal subtrees of T where position i appears to be conserved (that is, all sequences labeling a maximal subtree contain the same amino-acid at position i). For each i, it is easy to see that such decomposition of T into maximal subtrees is unique. For each such subtree T^i_j we evaluate the associated P^i_j. The computation of P^i_j is done as for P_0 above. Based on the P^i_js, for $j = 1 \dots S^i$, we compute the *evolution content at a position* i, denoted EC^i, with the

entropy function

$$EC^i = \gamma(n_i^\alpha \log_2 \beta n_i - \log_2 \beta)$$

where α, β, γ are parameters depending on the specific tree T we are working with. We define them and comment their significance below. The value n_i is computed from $n_i^* = \frac{\sum_{j=1}^{S^i} P_j^i}{P_0}$, where $\sum_{j=1}^{S^i} P_j^i \geq N$, by considering $\log_{10} n_i^*$ (this operation gives a value in the interval $[-x, 0]$, for some x) and by rescaling the result to the interval $[0, 1]$. The rationale is that the ratio $\frac{P_j^i}{P_0}$ represents the evolutionary distance between the root of T_j^i and the root of T. Larger the ratio is, closer the evolutionary content of the subtree is to the root. Note that for the leaves of the tree $P = 1$, since only one sequence is associated to a leaf.

The *evolution content of a tree of aligned sequences* is defined as

$$EC = \sum_{i=1}^{L} EC^i$$

To estimate the values of the parameters α, β, γ for a given protein we randomly select disjoint subtrees W_j in T in such a way that all leaves in T belong to some subtree W_j. Let m be the number of selected subtrees. After selection, we compute an expected value $n^{*exp} = \frac{\sum_{j=1}^{m} P_j}{P_0}$ and rescale it (by applying first $\log_{10}$) to $n^{exp} \in [0, 1]$. In practice, the random generation has to be repeated a sufficiently large number of times (about 100 times for instance) and the effective n^{exp} (to be used in the analysis) can be defined to be the average of the expected n^{exp}'s issued by each random selection. When a set of proteins is considered instead of a single protein, we compute the average of the n^{exp}s estimated for each protein. For different topologies of T, the value n^{exp} may vary, since it is directly associated to a distribution of subtrees in T.

The parameters α, β allow us to model information on residue conservation for a specific set of sequences and its associated tree. Namely, α, β are set so to preserve the convexity of the entropy function within $[0, 1]$ and in such a way that n^{exp} becomes the x-value where the entropy function takes its minimum. The parameter γ guarantees the y-values of the entropic function to fall into the same interval, in case several sequences are considered. The constant $\log_2 \beta$ guarantees the maximum of the entropic function to be at 0. The computation of the parameters is described in Materials and Methods.

COMPARISON OF EC AND IC ON RESIDUE POSITIONS

The advantage of using the measure of information EC instead of the more classical IC is shown, as a proof of principle, for a homodimeric D-amino acid aminotransferase protein. We test the hypothesis that the most "informative" residues of the protein structure (pdb:1daa) (Sugio et al. 1995) are those residues lying at the protein homodimeric interface. For this we evaluate the prediction of the protein interaction site based on the two notions. For each position i of the sequence alignment, we compute the corresponding EC^i and IC^i and we rank accordingly all residue positions from the most to the less informative. We then evaluate, by taking gradually larger sets of top ranked positions (coverage), whether these positions lie into the known interface site of the protein or not. We found that the EC^i notion ranks with much more precision the interface site, which intuitively defines the region where (functional) information resides. Particularly, EC^i detects especially well that signals of conservation are missing in the complementary region and makes prediction of the protein interface more sharp

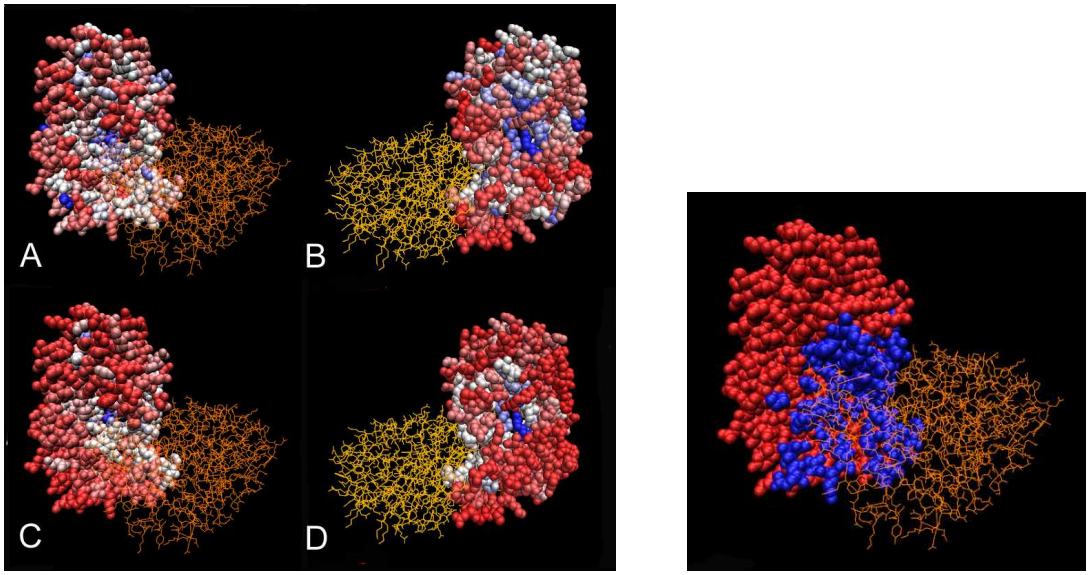

Figure 1: Left: A-B (top): two views of the protein where residue positions are colored with respect to their IC^i value. Red colors are associated to residues ranking low and blue colors to residues ranking high. The color scale starts at red, passes through rose and white to reach clear blue and blue. C-D (bottom): the same two views of the protein, as in A-B, where residue positions are colored with respect to their EC^i value. The color scale is the same as above. Note the rather sharply identifiable interaction site of the protein which is mainly colored white in C-D. In contrast, the scattering of white residues does not allow an easy identification of the interaction site in A-B. Right: residues belonging to the real interface are colored blue. All others are left red.

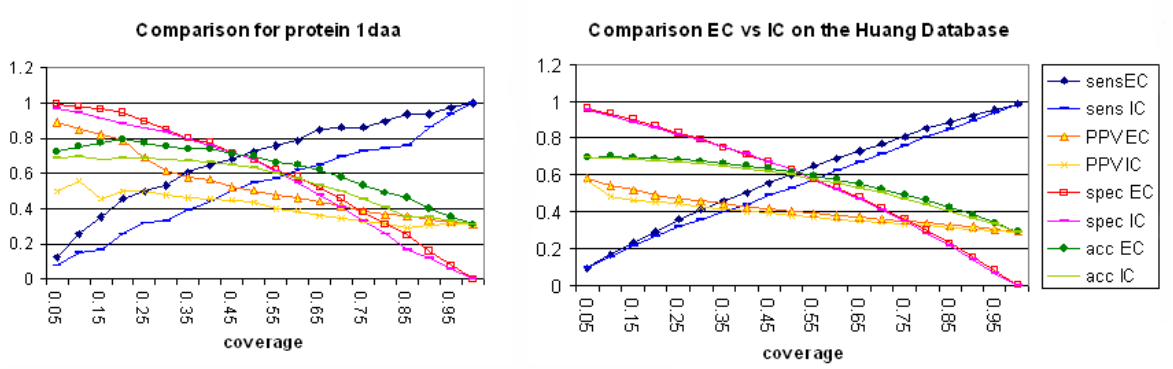

Figure 2: Left: Comparative evaluation of the predictions of the protein interaction site for the D-amino acid aminotransferase protein pdb:1daa based on the notions of EC and IC. Right: Comparative evaluation of the (average of the) predictions of the protein interaction site for all proteins in Huang Database based on the notions of EC and IC. The evaluation is realized on different coverage levels (x-axis).

(Figure 1). The numerical evaluation is reported for different coverages of the protein in the Appendix (Figure 3). See also Figure 2 (left).

The same analysis has being performed on a large database of 62 protein complexes, the Huang database (Caffrey et al. 2007), and for each protein complex the EC measure behaved better than the IC measure (with respect to all comparative scores). The analysis shows that homodimeric and heterodimeric protein interfaces gain in the EC evaluation, while transient protein interfaces are detected with sensitivity and PPV scores which are very low, that is close to random. See Figure 2 (right) for average evaluation scores computed on all protein complexes of the dataset, and see the Appendix for a numerical evaluation (Figure 4) and results on the three classes of protein interfaces (Figure 5).

MATERIALS AND METHODS

Protein complexes dataset for testing. The Huang database (Caffrey et al. 2007) of 62 protein complexes constituted by 41 homodimers (82 chains), 11 heterodimers (23 chains) and 8 transient complexes (17 chains) has been used to test the EC notion versus IC.

Evaluation. To properly compare the IC^i, EC^i notions on specific proteins, we rely on the following quantities: the number of residues correctly predicted as interacting (true positives, TP), the number of residues correctly predicted as non-interacting (true negatives, TN), the number of non-interacting residues incorrectly predicted as interacting (false positives, FP) and the number of interacting residues incorrectly predicted as non-interacting (false negatives, FN). We use four standard measures of performance: sensitivity $Sen = TP/(TP + FN)$, specificity $Spe = TN/(TN + FP)$, accuracy $Acc = (TP + TN)/(TP + FN + TN + FP)$ and positive predictive value $PPV = TP/(TP+FP)$. We also consider scores to evaluate the pertinence of the measures above with respect to expected values. Expected values are calculated on $TP^{exp} = C\cdot S, TN^{exp} = (1-C)(N-S), FP^{exp} = C\cdot(N-S), FN^{exp} = (1-C)\cdot S$, where $C = P/N$ is the coverage of the protein, where P is the number of surface residues predicted, N is the total number of surface residues and S is the number of residues in the real interaction site. Notice that the calculation of expected values assumes that $C \cdot N$ residues have been selected at random as being positives on the structure of the protein under study. This means that expected values are different for different proteins. Then we can compute sensitivity Sen^{exp}, specificity Spe^{exp}, accuracy Acc^{exp} and positive predictive value PPV^{exp} for the random case: $C, 1-C, ((1-C)\cdot(1-S/N)) + C\cdot S/N, S/N$ respectively. Pertinence scores are computed as follows: sensitivity score $ScSen = Sen - Sen^{exp}$, specificity score $ScSpe = Spe - Spe^{exp}$, accuracy score $ScAcc = Acc - Acc^{exp}$ and PPV score $ScPPV = PPV/PPV^{exp}$.

α, β, γ parameterisation for the entropy function applied to a single protein or to a dataset. The parameters α, β are set so to preserve the convexity of the entropy function within $[0, 1]$ and in such a way that the expected value n^{exp} for a given protein structure or a given database of proteins (defined above), becomes the minimum of the entropy function. Intuitively, while $\alpha \leq 1$ moves the minimum of the entropic function towards $n = 0$, the parameter β allows to start (at $\alpha = 1$) from a minimum which is close enough to n^{exp}.

The parameter β is the same for all sequences of a database. To explain its role let us consider its intrinsic relation with the parameter α. The equation $\log_2 \beta\cdot\alpha^2+(2-\log_2 \beta)\cdot\alpha-1 = 0$ expresses α in terms of β. There are two solutions α_1, α_2 for the equation and if α_1 falls into the interval $[0, 1]$, then the convexity of the function is guaranteed for $\alpha \in [\alpha_1, 1]$, otherwise it

is guaranteed for $(0, 1]$. By varying α within the interval $[\alpha_1, 1]$, the minimum of the entropic function falls into an interval I of the form $[\frac{e^{-1/\alpha_1}}{\beta}, \frac{e^{-1}}{\beta}]$. If α varies within $(0, 1]$ then the values of the entropic function fall within $I = (0, \frac{e^{-1}}{\beta}]$. By parameterizing β, we want all values n^{exp} associated to the full dataset of sequences (possibly one sequence) to belong to I. To do this, we fix β in such a way that $n \log_2 \beta n$ has the minimum at $\frac{e^{-1}}{\beta}$ (notice that here $\alpha = 1$).

For the Huang database, β takes value 2.1836, the interval of variation for α is $[0.5941, 1]$ and the minima of the entropic functions vary within $I = [0.0851, 0.1684]$. The interval I includes all values n^{exp} computed for the sequences of the Huang database.

To compare entropy values associated to trees of several different sequences, we use the parameter γ to guarantee the y-values of the entropic functions to fall into the same interval. Given α, β, we define γ to be $\frac{min\ (n \log_2 \beta n - \log_2 \beta)}{min\ (n^\alpha \log_2 \beta n - \log_2 \beta)}$. This way, all y-values of the entropic function fall into the common interval $[min\ (n \log_2 \beta n - \log_2 \beta), 0]$.

DISCUSSION

A numerical value coding for the information content of a structure is a very valuable quantity, and to correctly interpret this value is key for understanding how to use it. The classical definition of IC for a set of aligned amino-acid sequences is known to be representing the conservation level of the amino-acids in a protein. We show that it is only an "approximation" of this idea and that the conservation level can be described more properly by revisiting the IC notion with a new and very simple interpretation of the distance tree associated to multiple sequence alignment. The new notion EC provides a better estimation of the conserved interaction sites in a protein. Particularly, it detects especially well whether the complementary region of an interaction site is missing signals of conservation. This property is of particular importance when one wants to couple interaction site detection with docking algorithms. On a large database of protein complexes, we consistently observe that approximating IC with EC is always profitable.

One can envisage a definition of information content of sets of sequences that includes not only residue position conservation (coded in tree topology) but co-evolved residue positions, also coded in tree topology. This aim is far from being a trivial one. Notice that a similar attempt lead to the definition of information content for RNAs (Xayaphoummine et al. 2007), where RNA secondary structures provide a way to quantify feasible structures presenting co-evolving sites. For amino-acid sequences this task appears much more complicated due to the intrinsic physical-chemical nature of proteins. We should expect the new ranking induced by co-evolved residues to be correlated to sparse networks of amino-acids associated to functional and mechanical properties of the proteins in the sense of (Lockless and Ranganathan, 1999; Suel et al., 2003; Baussand 2008).

Some work related to our approach, is the study of spaces of sequences evolved for protein folding (Xia and Levitt 2004; Schmidt Am Busch et al. 2007). Also, an attempt to mix the notion of information content on sequences and the information coming from the tree topology has been proposed in (Mihalek et al. 2004). We demonstrated somewhere else that the definition reported there can be simplified by a better reading of the combinatorial structure of the tree (Engelen et al. 2008). In contrast to (Mihalek et al. 2004), notice that our contribution in this paper is to introduce a new explicit reading of distance trees from which to derive the information content of the pool of sequences.

ACKNOWLEDGMENTS

Part of this work has been done with the financial support of the AFM/IBM/CNRS Decrypthon project.

References

C. Adami , N.J. Cerf (2000). Physical complexity of symbolic sequences. *Physica D*, **137**, 62–69.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3 389–3402.

J. Baussand (2008) *Évolution des séquences protéiques: signatures structurales hydrophobes et réseaux d'acides aminés co-évolués.* Thèse de Doctorat de l'Université Pierre et Marie Curie-Paris 6.

D.R. Caffrey, S. Somaroo, J.H. Hughes, J. Mintseris, E.S. Huang (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, **13**, 190–189.

J.M. Carothers, S.C. Oestreich, J.H. Davis, J.W. Szostak (2004). Informational complexity and functional activity of RNA structures. *J Am Chem Soc.*, **126**, 5130–5137.

L. Duret, S. Abdeddaim (2000). Multiple alignment for structural functional or phylogenetic analyses of homologous sequences. In D. Higgins and W. Taylor: *Bioinformatics sequence structure and databanks.* Oxford: Oxford University Press.

S.Engelen, L.A. Trojan, S. Sacquin-Mora, R. Lavery, A. Carbone (2008). Joint Evolutionary Trees: detection and analysis of protein interfaces. Manuscript in preparation.

Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. (2001). Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene*, **270**, 17–30.

S. Lockless, R. Ranganathan (1999). Evolutionary conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.

I. Mihalek, I. Reš, O. Lichtarge (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.

J. Moult (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol.*, **15**, 285–289.

C. Notredame (2002). Recent progresses in multiple sequence alignment: a survey. *Pharmacogenomics*, **31**, 131–144.

C. Notredame (2007). Recent evolutions of multiple sequence alignment algorithms. *PLOS Computational Biology*, **8**, e123.

A. Phillips, D. Janies, W. Wheeler (2000) Multiple sequence alignment in phylogenetic analysis. *Mol Phylogenet Evol.*, **16**, 317–330.

M. Schmidt Am Busch, A. Lopes, D. Mignon, T. Simonson (2007). Computational protein design: Software implementation, parameter optimization, and performance of a simple model. *J Comput Chem.*, appeared online.

G. Suel, S. Lockless, M. Wall, R. Ranganthan (2003). Evolutionary conserved networks of residues mediate allosteric communication in proteins. *Nature Struct. Biol.*, **23**, 59–69.

S. Sugio, G.A. Petsko, J.M Manning, K. Soda, D. Ringe (1995). Crystal structure of a D-amino acid aminotransferase: how the protein controls stereoselectivity. *Biochemistry*, **34**, 9661–9669.

J.D. Thompson, F. Plewniak, O. Poch (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, **27**, 12682–12690.

I.M. Wallace, G. Blackshields, D.G. Higgins (2005). Multiple sequence alignments. *Curr Opin Struct Biol*, **15**, 261-266.

J.D. Watson, R.A. Laskowski, J.M. Thornton (2005). Predicting protein function from sequence and structural data. *Curr Opin Struct Biol.*, **15**, 275–284.

Xayaphoummine A, Viasnoff V, Harlepp S, Isambert H. (2007). Encoding folding paths of RNA switches. *Nucleic Acids Res.*, **35**, 614–622.

Y. Xia, M. Levitt (2004). Simulating Protein Evolution in Sequence and Structure Space. *Curr. Opin. Struct. Biol.*, **14**, 202–207.

Residues detected using EC^i in protein structure 1daa

cover	coverSurf	sens	ScoreSens	PPV	ScorePPV	spec	ScoreSpec	acc	ScoreAcc
0.0505	0.0424	0.1212	0.0787	0.8888	2.8551	0.9931	0.0355	0.7216	0.049
0.101	0.0943	0.2575	0.1632	0.85	2.7303	0.9794	0.0737	0.7547	0.1016
0.1516	0.132	0.3484	0.2164	0.8214	2.6385	0.9657	0.0978	0.7735	0.1347
0.2021	0.1792	0.4545	0.2753	0.7894	2.5358	0.9452	0.1244	0.7924	0.1714
0.2527	0.2264	0.5	0.2735	0.6875	2.2083	0.8972	0.1236	0.7735	0.1703
0.3032	0.2688	0.5303	0.2614	0.614	1.9723	0.8493	0.1181	0.75	0.1627
0.3501	0.3254	0.606	0.2805	0.5797	1.862	0.8013	0.1268	0.7405	0.1746
0.4007	0.3584	0.6515	0.293	0.5657	1.8173	0.7739	0.1324	0.7358	0.1824
0.4512	0.4056	0.6818	0.2761	0.5232	1.6807	0.7191	0.1248	0.7075	0.1719
0.5018	0.4481	0.7272	0.2791	0.5052	1.6229	0.678	0.1261	0.6933	0.1738
0.5523	0.4952	0.7575	0.2622	0.4761	1.5295	0.6232	0.1185	0.665	0.1633
0.6028	0.533	0.7878	0.2548	0.4601	1.4781	0.5821	0.1151	0.6462	0.1586
0.6534	0.5943	0.8484	0.2541	0.4444	1.4275	0.5205	0.1148	0.6226	0.1582
0.7003	0.6462	0.8636	0.2174	0.416	1.3364	0.452	0.0982	0.5801	0.1353
0.7509	0.6981	0.8636	0.1655	0.3851	1.237	0.3767	0.0748	0.5283	0.103
0.8014	0.7547	0.8939	0.1392	0.3687	1.1844	0.3082	0.0629	0.4905	0.0866
0.8519	0.8113	0.9393	0.128	0.3604	1.1578	0.2465	0.0578	0.4622	0.0797
0.9025	0.8726	0.9393	0.0667	0.3351	1.0764	0.1575	0.0301	0.4009	0.0415
0.953	0.9386	0.9696	0.031	0.3216	1.033	0.0753	1.40E-02	0.3537	0.0193
1	1	1	0	0.3113	0.9999	0	0	0.3113	0

Residues detected using IC^i in protein structure 1daa

cover	coverSurf	sens	ScoreSens	PPV	ScorePPV	spec	ScoreSpec	acc	ScoreAcc
0.0505	0.0471	0.0757	0.0285	0.5	1.606	0.9657	0.0129	0.6886	0.0177
0.101	0.0849	0.1515	0.0666	0.5555	1.7844	0.9452	0.0301	0.6981	0.0414
0.1516	0.1132	0.1666	0.0534	0.4583	1.4722	0.9109	0.0241	0.6792	0.0332
0.2021	0.1603	0.2575	0.0972	0.5	1.606	0.8835	0.0439	0.6886	0.0605
0.2527	0.1981	0.3181	0.12	0.5	1.606	0.8561	0.0542	0.6886	0.0747
0.3032	0.2169	0.3333	0.1163	0.4782	1.5362	0.8356	0.0525	0.6792	0.0724
0.3501	0.2641	0.3939	0.1297	0.4642	1.4913	0.7945	0.0586	0.6698	0.0808
0.4007	0.3018	0.4393	0.1375	0.4531	1.4554	0.7602	0.0621	0.6603	0.0856
0.4512	0.349	0.5	0.1509	0.4459	1.4324	0.7191	0.0682	0.6509	0.0939
0.5018	0.3915	0.5454	0.1539	0.4337	1.3931	0.678	0.0695	0.6367	0.0958
0.5523	0.4481	0.5757	0.1276	0.4	1.2848	0.6095	0.0576	0.599	0.0794
0.6028	0.5047	0.6212	0.1165	0.3831	1.2307	0.5479	0.0526	0.5707	0.0725
0.6534	0.566	0.6515	0.0854	0.3583	1.1509	0.4726	0.0386	0.5283	0.0532
0.7003	0.6273	0.6969	0.0696	0.3458	1.1109	0.4041	0.0314	0.4952	0.0433
0.7509	0.6886	0.7272	0.0385	0.3287	1.056	0.3287	0.0174	0.4528	0.024
0.8014	0.7452	0.7424	-0.0028	0.3101	0.9961	0.2534	-0.0012	0.4056	-0.0017
0.8519	0.8113	0.7575	-0.0537	0.2906	0.9337	0.1643	-0.0242	0.349	-0.0334
0.9025	0.8773	0.8636	-0.0137	0.3064	0.9843	0.1164	-0.0062	0.349	-0.0085
0.953	0.9433	0.9393	-0.004	0.31	0.9957	0.0547	-0.0018	0.3301	-0.0025
1	1	1	0	0.3113	0.9999	0	0	0.3113	0

Figure 3: Evaluation of the EC^i-ranking (top) and IC^i-ranking (bottom) on the D-amino acid amino-transferase protein structure pdb:1daa. Lines correspond to increasing coverage of the protein and describe prediction of the interaction site from 5% to 100% coverage.

Residues detected using EC^i in Huang database

cover	coverSurf	sens	ScoreSens	PPV	ScorePPV	spec	ScoreSpec	acc	ScoreAcc
0.0521	0.0432	0.0926	0.0494	0.5857	2.1094	0.9602	0.0034	0.699	0.0258
0.1021	0.0861	0.1677	0.0814	0.5437	1.922	0.9315	0.0177	0.7001	0.044
0.1518	0.127	0.2323	0.1052	0.5139	1.796	0.9008	0.0279	0.6975	0.0575
0.202	0.1697	0.2938	0.124	0.4921	1.7111	0.8668	0.0366	0.6918	0.0683
0.2516	0.2129	0.3547	0.1417	0.4736	1.6501	0.8302	0.0432	0.6841	0.0774
0.3019	0.2547	0.41	0.1552	0.4602	1.5962	0.7951	0.0499	0.6758	0.0854
0.352	0.2983	0.4585	0.1601	0.441	1.5207	0.7515	0.0499	0.6617	0.0884
0.4019	0.3421	0.5053	0.1631	0.4261	1.4617	0.7103	0.0526	0.6475	0.091
0.4519	0.3857	0.5543	0.1685	0.4156	1.4207	0.67	0.0558	0.634	0.0944
0.5009	0.4305	0.6037	0.1731	0.4055	1.3845	0.6276	0.0581	0.6184	0.0967
0.5521	0.4779	0.6465	0.1685	0.3922	1.3355	0.5792	0.0572	0.5972	0.0944
0.6017	0.5256	0.6873	0.1616	0.3807	1.2901	0.5295	0.0552	0.5745	0.0914
0.652	0.5753	0.7301	0.1547	0.3702	1.2521	0.4745	0.0499	0.5505	0.0877
0.7021	0.6268	0.7701	0.1432	0.359	1.212	0.4179	0.0447	0.5232	0.0812
0.7513	0.6814	0.8102	0.1288	0.3481	1.1726	0.3585	0.04	0.4933	0.0733
0.8017	0.7378	0.8534	0.1155	0.3389	1.14	0.2972	0.0352	0.4628	0.0659
0.8522	0.7979	0.8905	0.0925	0.3278	1.0997	0.2277	0.0256	0.4257	0.0532
0.9019	0.8589	0.9232	0.0642	0.3162	1.059	0.1539	0.0129	0.3843	0.0371
0.9519	0.9222	0.9567	0.0343	0.3054	1.0218	0.0775	-1.00E-04	0.3409	0.0198
1	0.9851	0.9851	0	0.2948	0.985	0	-0.0148	0.2948	0

Residues detected using IC^i in Huang database

cover	coverSurf	sens	ScoreSens	PPV	ScorePPV	spec	ScoreSpec	acc	ScoreAcc
0.0521	0.045	0.0911	0.046	0.5684	2.0399	0.9576	0.0027	0.6968	0.0243
0.102	0.0874	0.1529	0.0655	0.4862	1.7151	0.9232	0.0107	0.6907	0.035
0.1518	0.1283	0.2126	0.0842	0.4662	1.6318	0.89	0.0184	0.6849	0.0453
0.202	0.1672	0.2671	0.0998	0.4516	1.5824	0.8549	0.0222	0.6776	0.0533
0.2516	0.2046	0.3177	0.113	0.4413	1.5407	0.8233	0.028	0.6704	0.0606
0.3019	0.2405	0.3575	0.1169	0.4256	1.4772	0.79	0.0306	0.6597	0.0634
0.3521	0.2788	0.4011	0.1223	0.4152	1.425	0.7556	0.0345	0.6494	0.0678
0.4019	0.3187	0.44	0.1212	0.3995	1.3658	0.7163	0.0351	0.6339	0.0674
0.4519	0.3629	0.489	0.126	0.3892	1.3324	0.6709	0.0339	0.6181	0.0691
0.5009	0.4066	0.5319	0.1253	0.3788	1.2927	0.627	0.0337	0.6007	0.0689
0.5521	0.4569	0.5783	0.1213	0.3689	1.25	0.5771	0.0341	0.5798	0.0682
0.6017	0.5074	0.6245	0.117	0.36	1.2147	0.5267	0.0342	0.558	0.0667
0.652	0.5605	0.6714	0.1108	0.3503	1.1814	0.4716	0.0322	0.5321	0.0628
0.7021	0.6167	0.7182	0.1015	0.3422	1.1482	0.4134	0.0302	0.5055	0.059
0.7513	0.6751	0.7641	0.0889	0.3327	1.1157	0.3496	0.0248	0.4739	0.0515
0.8017	0.7355	0.8075	0.0719	0.3226	1.0819	0.2809	0.0166	0.4389	0.0411
0.8522	0.798	0.848	0.0499	0.3131	1.0469	0.2095	0.0077	0.4014	0.0293
0.9019	0.8605	0.8929	0.0323	0.3057	1.0221	0.1385	-8.00E-04	0.3651	0.0188
0.9519	0.9236	0.9386	0.0149	0.2998	1.001	0.0681	-0.0081	0.3296	0.009
1	0.9851	0.9851	0	0.2949	0.985	0	-0.0148	0.2949	0

Figure 4: Evaluation of the EC^i-ranking (top) and IC^i-ranking (bottom) on the Huang database. Lines correspond to increasing coverage of the protein and describe average predictions of the interaction site from 5% to 100% coverage computed for all 62 protein complexes of the database.

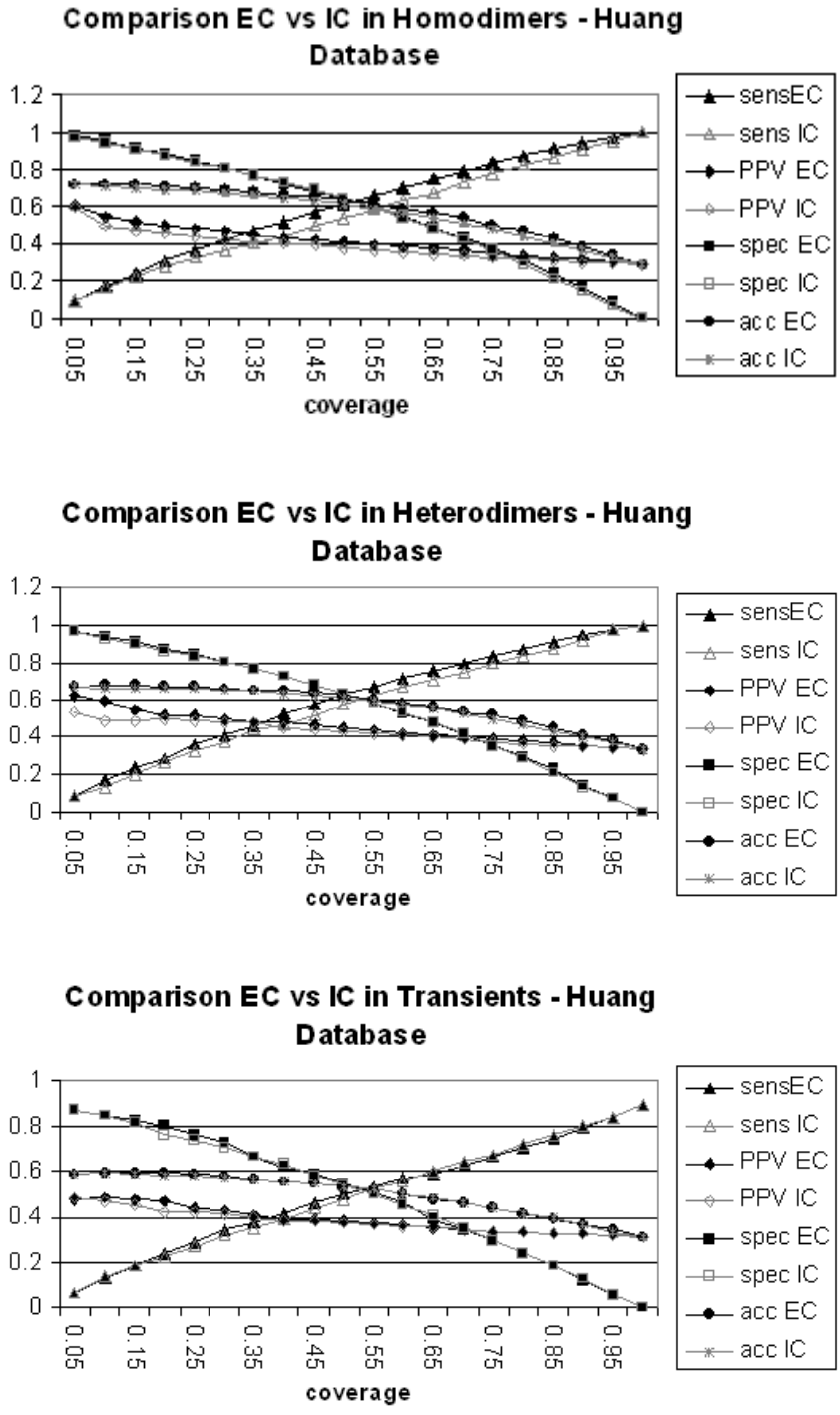

Figure 5: Comparative evaluation of the (average of the) predictions of homodimer (top), heterodimer (center) and transient (bottom) interfaces based on the notions of EC and IC. The evaluation is realized on different coverage levels.