Using orthologous and paralogous proteins to identify specificity determining residues

Deposited research article

Using orthologous and paralogous proteins to identify specificity determining residues

Leonid A. Mirny* and Mikhail S. Gelfand‡

Addresses: *Harvard-MIT Division of Health Science and Technology, Massachusetts Institute of Technology, 77 Massachusetts ave, Cambridge, MA 02139.‡IntegratedGenomics-Moscow, P.O.Box 348, Moscow, 117333, Russia.

Correspondence: Leonid A. Mirny. E-mail: leonid@mit.edu

Posted: 19 February 2002

Genome ***Biology*** 2002, **3(3)**:preprint0002.1-0002.20

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/3/preprint/0002

Received: 13 February 2002

This is the first version of this article to be made available publicly.

Using orthologous and paralogous proteins to identify specificity determining residues.

Leonid A. Mirny* and Mikhail S. Gelfand‡

∗ Corresponding author: Harvard-MIT Division of Health Science and Technology, Massachusetts Institute of Technology, 77 Massachusetts ave, Cambridge, MA02139, leonid@mit.edu

‡ IntegratedGenomics-Moscow, P.O.Box 348, Moscow, 117333, Russia

Abstract

Background Concepts of orthology and paralogy are become increasingly important as whole-genome comparison allows their identification in complete genomes. Functional specificity of proteins is assumed to be conserved among orthologs and is different among paralogs. We used this assumption to identify residues which determine specificity of protein-DNA and protein-ligand recognition. Finding such residues is crucial for understanding mechanisms of molecular recognition and for rational protein and drug design.

Results Assuming conservation of specificity among orthologs and different specificity of paralogs, we identify residues which correlate with this grouping by specificity. The method is taking advantage of complete genomes to find multiple orthologs and paralogs. The central part of this method is a procedure to compute statistical significance of the predictions. The procedure is based on a simple statistical model of protein evolution. When applied to a large family of bacterial transcription factors, our method identified 12 residues that are presumed to determine the protein-DNA and protein-ligand recognition specificity. Structural analysis of the proteins and available experimental results strongly support our predictions. Our results suggest new experiments aimed at rational re-design of specificity in bacterial transcription factors by a minimal number of mutations.

Conclusions While sets of orthologous and paralogous proteins can be easily derived from complete genomic sequences, our method can identify putative specificity determinants in such proteins.

Background

The concepts of orthology and paralogy were originally introduced by Walter Fitch in 1970 [1, 2] and recently became a subject of active discussion in Genome Biology [3, 4, 5, 6]. Briefly, orthologs are genes in different organisms which are direct evolutionary counterparts of each other. Orthologs were inherited through speciation, as opposed to paralogs which are genes in the same organism which evolved by gene duplication [6, 3, 2]. After duplication, paralogous proteins experience weaker evolutionary pressure and their specificity diverges leading to emerging of new specificities and functions. Orthologous proteins, on the contrary, are believed to be under similar regulation, have the same function and usually the same specificity in close organisms [7, 8, 9]. In other words, both paralogs and orthologs are assumed to have similar general biochemical functions, while orthologs are also believed to have the same specificity. Although validity of these assumptions are yet to be verified experimentally, numerous case studies support such views [6, 10]. Several methods have been developed to find orthologous proteins in complete genomes [8, 11]. The assumption of similar regulation of orthologous proteins was productively used by several groups to identify common regulatory motifs upstream of orthologous proteins [12, 9, 13, 14, 15]. In this study we exploit another property of orthologs - similar specificity, as contrasted by different specificities of paralogs .

If the assumption above is correct, grouping by orthology becomes grouping of proteins by specificity. Here we developed a method which uses such grouping to identify amino acids which determine the protein specificity. Specificity-determining residues can be very hard to find even when the structure of a protein or a complex is available, since very few amino acids provide specific recognition (see below). Extensive site-directed mutagenesis is used to find such residues, though frequently complicated by a need to discriminate between specific and non-specific effects of a mutation. Computational prediction of the specificity determinants can substantially reduce experimental efforts and provide guidance for rational re-design of protein function [16, 17].

Our method relies on the above assumption that binding specificity is conserved among orthologous proteins and is different in paralogous proteins. The idea of our method is (1) to start from a family of paralogs in one genome, find orthologs for each member of the family in other genomes and (2) identify residues that can better discriminate between these orthologous (specificity) groups.

In its second part the method is similar to techniques of Hierarchical Analysis of Residue Conservation [18], PCA in the sequence space [19], Evolutionary Trace Analysis [20, 21] and Prediction of Functional Sub-types [22]. All these techniques use multiple sequence alignment to group proteins into sub-groups based on sequence

similarity and then identify residues that confer the unique features of each subgroup. A complementary structure-based approach was developed by Johnson and Church to predict protein function using a prior knowledge of the binding-site residues [23]. In contrast to other methods, ours relies on definition of sub-families based on gene orthology and a rigorous statistical procedure to predict specificity determining residues. Our statistical procedure determines whether positions in the MSA can discriminate between functional sub-families better than the sequence similarity. Residues that satisfy these criteria are predicted to be specificity-determining. Importantly, out method does not require the knowledge of the protein structure and can tolerate certain substitutions within a sub-family.

Here we present results of our analysis applied to the LacI/PurR family of bacterial transcription factors. The main result of this study is that among 12 identified specificity determining residues, three are binding the DNA and eight are binding the ligand in the ligand-binding domain. The available experimental information supports the critical role of the identified DNA-binding residues in determining the specificity of the DNA recognition. Analysis developed here is not limited to DNA-binding proteins and can be applied to any family of proteins where the clear orthology or functional grouping can be established.

Methods

The key idea of this method is to compare paralogous and orthologous proteins from the same family. As a rule, all paralogous and orthologous proteins have the same biochemical function. Paralogous proteins, however, usually have different specificity as they act on different targets, e.g. bind different ligand or different sites on the DNA. Orthologous proteins, in contrast, have the same specificity in different organisms, e.g. bind the same ligand and similar DNA sites in related genomes. Hence, orthologous proteins carry the same or similar specificity determining residues, whereas paralogous proteins carry different ones. Based on this idea, our analysis is looking for residues that are conserved among orthologs and different in paralogs. More generally, we are looking for residues that can *discriminate* between different paralogs, while grouping orthologs together. We call these residues *specificity determining*.

The analysis works as following: First, in a group of homologous proteins, paralogs from one organism are selected. Second, for each of the paralogs we find its orthologs in related organisms and build a multiple sequence alignment (MSA) using ClustalW [24]. Third, we compute the mutual information for each position of the MSA. The mutual information determines how well a residue in the MSA can discriminate between orthologous groups. The forth, and the most important step is to

compute the *statistical significance* of the discrimination and to select residues that can discriminate significantly better than the others. These residues are the specificity determinants.

Selection of orthologs

A list of complete and almost complete bacterial genomes used in this study and a full list of orthologs is provided in Supplementary Information. Homologs of LacI and PurR of *E. coli* were identified using GenomeExplorer [25] and supplemented by proteins from SwissProt [26]. Then Phylogenetic trees were constructed using the neighbor-joining procedure implemented in ClustalW [24]. Only unambiguous groups of orthologs identified by (1) absence of duplications in corresponding sub-branches of the tree (in two cases duplications in one genome were allowed where the proteins were known to have the same ligand and DNA-binding specificity), (2) functional annotation when known, and (3) genomic positional analysis [11] were selected.

Mutual information

The goal is to identify residues that can discriminate between paralogous proteins (different specificity) and at the same time merge orthologs (same specificity) together. To find such residues we use the mutual information as a measure of association with the specificity. Mutual information is frequently used in computational biology for co-variational analysis in RNA and proteins [27, 28].

If $x = 1 \dots 20$ is a residue type, $y = 1 \dots Y$ is the specificity index which is the same for all proteins of the same specificity group and is different for different groups, and Y is the total number of specificity groups, then the mutual information at position i of the MSA is:

$$I_i = \sum_{\substack{x=1\dots20 \\ y=1\dots Y}} f_i(x,y) \log \frac{f_i(x,y)}{f_i(x) f(y)} \qquad (1)$$

where $f_i(x)$ is the frequency of residue type x in position i of the MSA, $f(y)$ is the fraction of proteins belonging to the group y, and $f_i(x,y)$ is the frequency of residue type x in the group y at position i. Mutual information has several important properties: (1) it is nonnegative; (2) it equals zero if and only if x and y are statistically independent; and (3) a large value of I_i indicates a strong association between x and y [29]. Unfortunately, a small sample size and a biased composition of each column in the MSA influences I_i a lot. For example, positions with less conserved residues tend to have higher mutual information. Hence, we can not rely on the value of I_i as an indicator of specificity association, instead we estimate the statistical significance of I_i.

Statistical Significance

Since mutual information can be biased due to the small sample size or biased amino acid composition, we can not rely on the value of mutual information to identify the specificity determinants. Instead, we compute the statistical significance $P(I)$ of the mutual information and use it together with I to predict the specificity determining residues. Calculation of statistical significance is the most important component of the method. We present two different approaches, which, however, produce very similar results.

A standard way of computing $P(I)$ is by shuffling [30]. However, this method is unacceptable for the following reason. Naturally, proteins within each specificity group (orthologs) are much more similar to each other than proteins from different groups (paralogs). Hence, amino acids at *every* position are somewhat associated with the functional grouping, producing I higher than the mutual information obtained by shuffling (data not shown). Developing a statistical test we have to take into account the naturally higher similarity between orthologs in comparison to paralogs. In other words, we need a statistical test to *identify positions that are stronger associated with the functional grouping, than the whole proteins on average.*

To accomplish this task we developed two statistical models. The first model uses linear transformation to take into account a bias introduced by higher intra-group similarity of orthologs. The second model simulates accumulation of mutations in duplicated genes in order to reproduces a higher intra-group similarity. Details of both statistical models are given in Supplementary Information.

Briefly, in the first model we begin with computing the distribution of I^{sh} by multiple shuffling of every position in the MSA. Then the distribution of I^{sh} is transformed into the distribution of expected mutual information $I_i^{exp} = \alpha I_i^{sh} + \beta$. Parameters α and β are set to have the same value for all positions i. The values of α and β are found by maximal likelihood estimator. This transformation compensates for position-independent bias due to higher intra-group similarity of orthologs. Distribution of I^{exp} is then used to compute sought statistical significance $P_i(I)$.

The second model does not use shuffling. Instead, we simulate evolution of proteins in the family and generate a set of pseudo-random protein sequences. The proteins are generated such that (i) composition of each column in the MSA of the pseudo-random proteins is the same as in the real proteins; and (ii) intra- and inter-group similarity resembles those of the real proteins. Next, we compute the distribution of mutual information I^{rnd} for the MSA of pseudo-random proteins. This distribution is used to computer $P_i = P(I_i)$. See Supplementary Information for details.

Results

Specificity determinants of the LacI family

We have chosen the LacI family for our analysis because (1) it is one of the largest families of bacterial transcription factors, (2) the availability of complete bacterial genomes has allowed us to resolve orthology by positional analysis (see Methods), and (3) available experimental [31, 32, 33] and structural [34, 35] information can be used to verify our predictions.

Figure 1 presents the mutual information I_i, the expected mutual information I^{exp} and the probability $P(I)$ computed for the LacI family using Model1. Model2 produces very similar results (see Supplementary Information). This plot reveals several important features: First, it shows high correlation $\rho = 0.97$ between I_i and I_i^{exp}. Very good agreement between I_i and I_i^{exp} demonstrates that statistical model used to compute I^{exp} succeeded in explaining $\rho^2 = 94\%$ of variation in mutual information and is able to reproduce naturally higher mutual information due to high intra-family similarity of orthologs. Second, the vast majority of amino acids in the LacI family exhibit weak association with the specificity as indicated by $P(I) \approx 1$. Third, very few positions have both low $P(I_i)$ and high I_i (shown by arrows on Fig 1). Amino acids in these positions have strong association with functional grouping (stronger than sequences on average), indicating the role of these positions in determining different specificities of different groups of orthologs.

Table 1 presents predicted specificity determining amino acids. Importantly, although methods to estimate statistical significance are very different, sets of residues found by them are very similar. The specificity determinants are: 15, 16, 50 and 55, in the first domain; and 98, 114, 122, 146, 147, 160, 221 and 249 in the second domain (here and below the numbering is according to PurR; the PDB code 1wet).

Table 2 of Supplementary Information shows the pattern of conservation of predicted specificity determinants. As expected, most of these residues are conserved within orthologous groups and are different between different groups. Importantly, there are some of exceptions from this rule in all specificity determining positions (see Discussions).

To better understand the role of specificity determining residues we map them onto the structures of the PurR and LacI-DNA complexes. Figure 2 presents the structure of the PurR-DNA complex with specificity-determining residues shown by space-filling atomic models with atoms of van der Waals radii. Clearly, these residues form two clusters in the structure: one around the DNA and another around the ligand. This result comes at no surprise, since proteins of the LacI family act as transcription

repressors (activators) upon presence or absence of particular small molecules (sugars, nucleotides etc). Hence, paralogous proteins differ in specificity of both DNA and small molecule (ligand) recognition. The two identified spatial clusters supposedly determine this specificity.

Examination of the structure brings us to the following conclusions. (1) First four specificity-determining residues in PurR THR15, THR16, VAL50 and LYS55 (TYR17, GLN18, VAL52 and ALA57 in LacI) are located in the DNA-binding domain. Three of them (15, 16 and 55 in PurR; 17,18,57 in LacI) are deeply buried in the DNA grooves forming a dense network of interactions with the bases (see Fig. 3C,D). VAL50 (VAL52 in LacI) forms a hydrophobic contact with its counterpart on the other chain. (2) Six more specificity-determining residues (out of eight) MET122, ASP146, TRP147, ASP160, PHE221, ILE249 (ASN125, ASP149, VAL150, PHE161, TRP220, GLN248 in LacI) are located in the ligand-binding pocket. Five of them (MET122, ASP146, ASP160, PHE221, ILE249) are within $8\mathring{A}$ from the ligand in PurR and within $5\mathring{A}$ in LacI (ASN125, ASP149, PHE161, TRP220, GLN248) (see Fig. 3A,B). The observed clustering of the identified amino acids around the ligand is striking since the structure of the protein was not used in our analysis.

Such structural location indicates that identified residues are indeed involved in the specific recognition. While the DNA-binding residues determine motifs recognized on the DNA, the residues located close to the ligand determine the ligand-binding specificity of the protein. Since different orthologs have different ligands, these residues change from sub-family to sub-family, but stay the same within most sub-families. PHE221 in PurR and corresponding TRP220 in LacI are of a special interest as their aromatic rings directly interact with aromatic ligands. Two other residues, (TRP98 and LYS114 in PurR; ARG101, GLN117 in LacI) do not belong to either of the clusters, as they are located far from the DNA and the ligand. They either are "false positives", or have some special role in the alosteric regulation [36]. Indeed, VAL50, TRP98 and LYS114 of one chain interact tightly with the other chain, specificly VAL50 interacts with LYS114 of the other chain. These residues can be important for correct dimerization and hence exhibit sought covariation with functional grouping. In summary, the structural location of identified residues supports the view that they serve as specificity determinants in proteins of the LacI family. This includes the specificity of the DNA recognition and the ligand-binding specificity.

Discussion

In this study we suggested a method to identify specificity determining residues in proteins. We applied it to one of the largest family of bacterial transcription factors

and obtained a set of putative specificity determining residues. Mapping of these residues onto a protein structure showed that most of identified residues belong to two spatial clusters. Residues of one cluster bind the DNA, while residues of the other cluster form a ligand pocket of the protein. This finding is consistent with the function of transcription factors of this family: they repress transcription by binding the DNA and release transcription when a particular ligand is present. (Conversely, some proteins in the family, e.g. PurR, bind the DNA only when the ligand is present). Paralogous proteins of this family differ from each other in the ligands they recognize and in the DNA sites they bind. Hence, two clusters of residues found by our method presumably determine specificity of these two recognition processes.

Our analysis suggested residues 15, 16 and 55 as primary determinants of the DNA-binding specificity. The role of positions 15, 16 and 55 in specific DNA recognition is evident from a series of mutant experiments [31, 32, 33]. Extensive site-directed mutagenesis of the second helix of LacI showed that residues 15 and 16 are essential for DNA-binding specificity [37]. When TYR15 and GLN16 of LacI were mutated to the residues present in these positions in the paralogs (malI, rafR, cytR etc) the mutants were preferentially binding operators of the respective paralogs. Similarly, when GalR was mutated to have the residues of LacI in positions 15 and 16, mutant GalR was specifically binding sites of LacI. These experiments strongly support our result that positions 15 and 16 are responsible for determining DNA-binding specificity in proteins of the LacI family. Another residue found by our analysis is residue 55. Although 55 is binding DNA in the minor groove, this residue was shown to be critical for the DNA recognition by PurR [33]. Our results suggest that a triple mutant (15,16 and 55) should have a higher specificity and affinity to paralogous operators.

To the best of our knowledge, residues identified in the ligand-binding domain (except for 146) have not yet been the subjects of protein engineering studies. Although mutations of several other residues were shown to interfere with the ligand binding, it is not clear how they influence specificity (as opposed to affinity) of the ligand recognition. Most of mutations in the region were shown to drastically reduce the affinity. Our analysis suggests ways to do rational re-design of the ligand binding specificity. One can "transplant" some or all of the outlined residues from a paralog to LacI and measure the mutant's binding constants for various ligands normally bound by this paralog. The main question posed by our study is whether the specificity can be re-designed by changing a small set of the predicted residues.

Another possible application of the putative specificity determinants is in more focused prediction of the DNA-binding specificity. Instead of considering all interactions between the DNA and the protein, one can focus on the interactions formed by the

specificity determinants. This approach is a subject of our current research.

The most important part of presented algorithm is the procedure used to calculate statistical significance of the mutual information. Specificity determinants were selected as residues having both high I and very low $P(I)$. Note, that selection by high I alone would yield a very different (and very large) set of residues (see Fig. 1, filled circles). Most of such residues do not have statistically significant association with grouping ($P(I) \approx 1$). This observation emphasizes the importance of statistical test in our analysis.

Although promising, our analysis has its limitations. It relies heavily on the grouping of proteins by orthology. To resolve orthology, one needs to have (almost) complete genomes of several closely related organisms. This makes our analysis significantly data demanding. Even if complete genomes are available, orthology may not be easily resolved when very similar paralogs are present or when genomes are too diverged from each other. Our analysis also assumes that orthologs have the same function and specificity. This is likely true for evolutionary close organisms, where orthologs had not enough evolutionary time to diverge in specificity. One way to avoid these pitfalls is to use proteins where conserved specificity has been experimentally verified or confirmed by independent genomic positional or regulatory analysis. Using genomically resolved orthologs one has to rely on the statistical significance, $P(I)$. If a dataset is "contaminated" with false orthologs or orthologs with diverged specificity, no residues would have low $P(I)$.

Our analysis also relies on the assumption that the same residues determine specificity of paralogs. Little is known about the spatial location of the specificity determinants. Active site residues, however, are known to have very conserved spatial location in the families of homologous proteins. Active sites have the same spatial location, even when similarity between the sequences is as low as 10%. For example, proteins of the TIM-barrel and flavodoxin folds have "super-sites", i.e. active sites in the same spatial location, although amino acids forming the sites and the biochemical function of these proteins have widely diverged [38, 39]. This extreme conservation of the spatial location of functionally crucial residues supports the assumption of common location of the specificity determinants. Since most orthologs and close paralogs have high sequence similarity, residues matched in the multiple sequence alignments are likely to have common spatial location. However, recognition of molecules of various shapes (ligands, protein interfaces) may involve different interactions and therefore variable spatial location of the specificity residues can not be ruled out. The most direct test of our assumption would be to perform the experiments suggested above. Calculation of statistical significance also constitutes an internal test of our method. If sets of residues which determine specificities of paralogs differ, our method will

identify an overlap between these sets as having low $P(I)$. If the sets do not overlap, no residues would have substantially low $P(I)$.

As can be seen from Table 2 of Supplementary Information our method, in contrast to other methods, can tolerate certain substitutions within a group of orthologs. All amino acids, however, are assumed to be equally distinct in their properties. In other words, substitutions $I \rightarrow L$ and $I \rightarrow H$ are treated equally, while in reality the change of the physical properties of amino acids depends on the type of the substitutions. We are currently developing a method to identify specificity determinants with different spatial locations, which will also take into account physical properties of amino acids.

Here we have suggested a method to find residues that determine the specificity of the protein recognition. The method is based on discrimination between orthologous and paralogous proteins, taking advantage of several complete bacterial genomes to identify them. The method does not require a solved 3D structure of a protein to predict specificity determinants. Analysis of a large LacI family of bacterial transcription factors found two groups of residues as the putative DNA-binding and ligand-binding specificity determinants. Predictions of the DNA-binding residues are strongly supported by the earlier experimental results. Results of our analysis suggest targeted protein engineering experiments aimed at rational re-design of the protein specificity.

An additional file available with the online version of this article includes detailed description of the methods used to compute statistical significance and a table of specificity determining amino acids in other proteins of LacI family.

We are grateful to Alexander van Oudenaarden for helpful discussions and initiation of experimental work to test our predictions. We also acknowledge useful comments made by Richard Goldstein and Eugene Shakhnovich while discussing this work. LM is partially supported by William F. Milton Fund and John F. and Virginia B. Taplin Award. MG is partially supported by grants from INTAS (99-1476), Howard Hughes Medical Institute (55000309), and the Ludwig Cancer Research Institute. We are grateful to Dmitry Rodionov for the help with the data and Olga Laikova for useful discussions.

References

1. Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19**:99–113.

2. Fitch WM: **Homology a personal view on some of the problems.** *Trends Genet* 2000, **16**:227–31.

3. Koonin EV: **An apology for orthologs - or brave new memes.** *Genome Biol* 2001, **2**:1005.

4. Petsko GA: **Homologuephobia.** *Genome Biol* 2001, **2**:1002.

5. Jensen RA: **Orthologs and paralogs - we need to get it right.** *Genome Biol* 2001, **2**:1002.

6. Gerlt JA, Babbitt PC: **Can sequence determine function?** *Genome Biol* 2000, **1**:5.

7. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV: **Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res* 1999, **9**:608–28.

8. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28**:33–6.

9. Gelfand MS, Koonin EV, Mironov AA: **Prediction of transcription regulatory sites in archaea by a comparative genomic approach.** *Nucleic Acids Res* 2000, **28**:695–705.

10. Gerlt JA, Babbitt PC: **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies.** *Annu Rev Biochem* 2001, **70**:209–46.

11. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci U S A* 1999, **96**:2896–901.

12. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia Coli* K-12.** *Nucleic Acids Res* 2001, **29**:72–4.

13. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29**:774–82.

14. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**:563–77.

15. Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD: **A comparative genomics approach to prediction of new members of regulons.** *Genome Res* 2001, **11**:566–84.

16. Fersht AR: *Structure and Mechanism in Protein Science : A Guide to Enzyme. Catalysis and Protein Folding.*: WH Freeman & Co, San Francisco, 1999.

17. Ballinger MD, Tom J, Wells JA: **Furilisin: a variant of subtilisin bpn' engineered for cleaving tribasic substrates.** *Biochemistry* 1996, **35**:13579–85.

18. Livingstone CD, Barton GJ: **Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.** *Comput Appl Biosci* 1993, **9**:745–56.

19. Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins.** *Nat Struct Biol* 1995, **2**:171–8.

20. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257**:342–58.

21. Lichtarge O, Yamamoto KR, Cohen FE: **Identification of functional surfaces of the zinc binding domains of intracellular receptors.** *J Mol Biol* 1997, **274**:325–37.

22. Hannenhalli SS Russell RB: **Analysis and prediction of functional sub-types from protein sequence alignments.** *J Mol Biol* 2000, **303**:61–76.

23. Johnson JM Church GM: **Predicting ligand-binding function in families of bacterial receptors.** *Proc Natl Acad Sci U S A* 2000, **97**:3965–70.

24. Thompson JD, Higgins DG, Gibson TJ: **Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties.** *Nucleic Acids Res* 1994, **22**:4673–80.

25. Mironov AA, Vinokurova NP, Gelfand MS: **Software for analysis of bacterial genomes** *Mol Biol (Mosk)* 2000, **34**:253–62.

26. Bairoch A Apweiler R: **The Swiss-Prot protein sequence database and its supplement TREMBL in 2000.** *Nucleic Acids Res* 2000, **28**:45–8.

27. Clarke ND: **Covariation of residues in the homeodomain sequence family.** *Protein Sci* 1995, **4**:2269–78.

28. Gorodkin J, Staerfeldt HH, Lund O, Brunak S: **Matrixplot: visualizing sequence constraints.** *Bioinformatics* 1999, **15**:769–70.

29. Cover TM ,Thomas JA: *Elements of information theory.* Wiley, New York, 1991.

30. Good PI: *Permutation tests : a practical guide to resampling methods for. testing hypotheses* Springer series in statistics Springer-Verlag, New York, 1994.

31. Sartorius J, Lehming N, Kisters-Woike B, von Wilcken-Bergmann, Muller-Hill B: **The roles of residues 5 and 9 of the recognition helix of Lac repressor in Lac operator binding.**: *J Mol Biol* 1991, **218**:313–21.

32. Lehming N, Sartorius J, Kisters-Woike B, von Wilcken-Bergmann, Muller-Hill B: **Mutant lac repressors with new specificities hint at rules for protein–DNA recognition.** *EMBO J* 1990, **9**:615–21.

33. Glasfeld A, Koehler AN, Schumacher MA, Brennan RG: **The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions.** *J Mol Biol* 1999, **291**:347–61.

34. Schumacher MA, Glasfeld A, Zalkin H, Brennan RG: **The X-ray structure of the PurR-guanine-PurF operator complex reveals the contributions of complementary electrostatic surfaces.** *J Biol Chem* 1997, **272**:22648–53.

35. Bell CE, Lewis M: **Crystallographic analysis of Lac repressor bound to natural operator O1.**: *J Mol Biol* 2001, **312**:921–6.

36. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286**:295–9.

37. Sartorius J, Lehming N, Kisters B, von Wilcken-Bergmann, Muller-Hill B: **lac repressor mutants with double or triple exchanges in the recognition helix bind specifically to lac operator variants with multiple exchanges.** *EMBO J* 1989, **8**:1265–70.

38. Hasson MS, Schlichting I, Moulai J, Taylor K, Barrett W, Kenyon GL, Babbitt PC, Gerlt JA, Petsko GA, Ringe D: **Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase.** *Proc Natl Acad Sci U S A* 1998, **95**:10396–401.

39. Russell RB, Sasieni PD, Sternberg MJ: **Supersites within superfolds. binding site similarity in the absence of homology.** *J Mol Biol* 1998, **282**:903–18.

40. Kunst F, Ogasawara N, Moszer I, et.al: **The complete genome sequence of the gram-positive bacterium *Bacillus Subtilis*.** *Nature* 1997, **390**:249–56.

41. Nolling J, Breton G, Omelchenko MV et.al: **Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium Acetobutylicum*.**: *J Bacteriol* 2001, **183**:4823–38.

42. Ferretti JJ, McShan WM, Ajdic D et al: **Complete genome sequence of an M1 strain of *Streptococcus pyogenes*.** *Proc Natl Acad Sci U S A* 2001, **98**:4658–63.

43. Tettelin H, Nelson KE, Paulsen IT et al: **Complete genome sequence of a virulent isolate of *Streptococcus Pneumoniae*.** *Science* 2001, **293**:498–506.

44. Heidelberg JF, Eisen JA, Nelson WC et al: **DNA sequence of both chromosomes of the cholera pathogen *Vibrio Cholerae*.** *Nature* 2000, **406**:477–83.

45. Blattner FR, Plunkett G, Bloch CA et al: **The complete genome sequence of *Escherichia Coli* K-12.** *Science* 1997, **277**:1453–74.

46. Parkhill J, Wren BW, Thomson NR et al: **Genome sequence of *Yersinia Pestis*, the causative agent of plague.** *Nature* 2001, **413**:523–7.

47. Fleischmann RD, Adams MD, White O et al: **Whole-genome random sequencing and assembly of *Haemophilus Influenzae*.** *Science* 1995, **269**:496–512.

48. Stover CK, Pham XQ, Erwin AL at al: **Complete genome sequence of *Pseudomonas Aeruginosa* PA01, an opportunistic pathogen.** *Nature* 2000, **406**:959–64.

Figures and Tables

Figure 1 Observed I (blue) and the mean expected I^{exp} (thick red) mutual information in DNA-binding (A) and ligand-binding (B) domains of LacI family. Thin red lines show $I^{exp} \pm 2\sigma(I^{exp})$. $P(I)$ is statistical significance of mutual information. Filled circles indicated residues with $I > 1.0$. Positions with filled circles and low $P(I)$ are predicted specificity determinants. The number along the sequence are according to 1wet PDB structure.

Figure 2 Structure of PurR bound to the DNA. Two chains of the dimer are shown semi-transparent in light green and pink. Predicted specificity determinants are shown by space-filling and colored red in the pink chain and green in the light green chain. The ligand () and the DNA are shown in blue. Notice deep penetration of some specificity-determining residues into the DNA and formation of the ligand-binding pocket by most of the others.

Figure 3 Detailed picture of the ligand binding pockets (A,B) and protein-DNA interface (C,D) in PurR (left) LacI (right). Predicted specificity determinants are shown in space-fill.

Table 1 Lists of specificity-determining residues as predicted by different methods. Numbering is according to 1wet PDB file. See Supplementary Information for equations and details.

Table 1:

Method		DNA-binding Domain		Ligand-binding Domain
Model 1, eq (3)	$P < 10^{-5}$ $I > 1.0$:	15,16,50,55	$P < 10^{-5}$ $I > 1.5$:	98,114,122,146,147,160,221,249
Model 1, eq (4)	$P < 10^{-4}$ $I > 1.0$:	15,16,50,55	$P < 10^{-4}$ $I > 1.5$:	98,114,146,160,221
Model 2	$P < 10^{-2}$:	15,16,55	$P < 10^{-2}$:	85,98,122,146,160,221,246,249

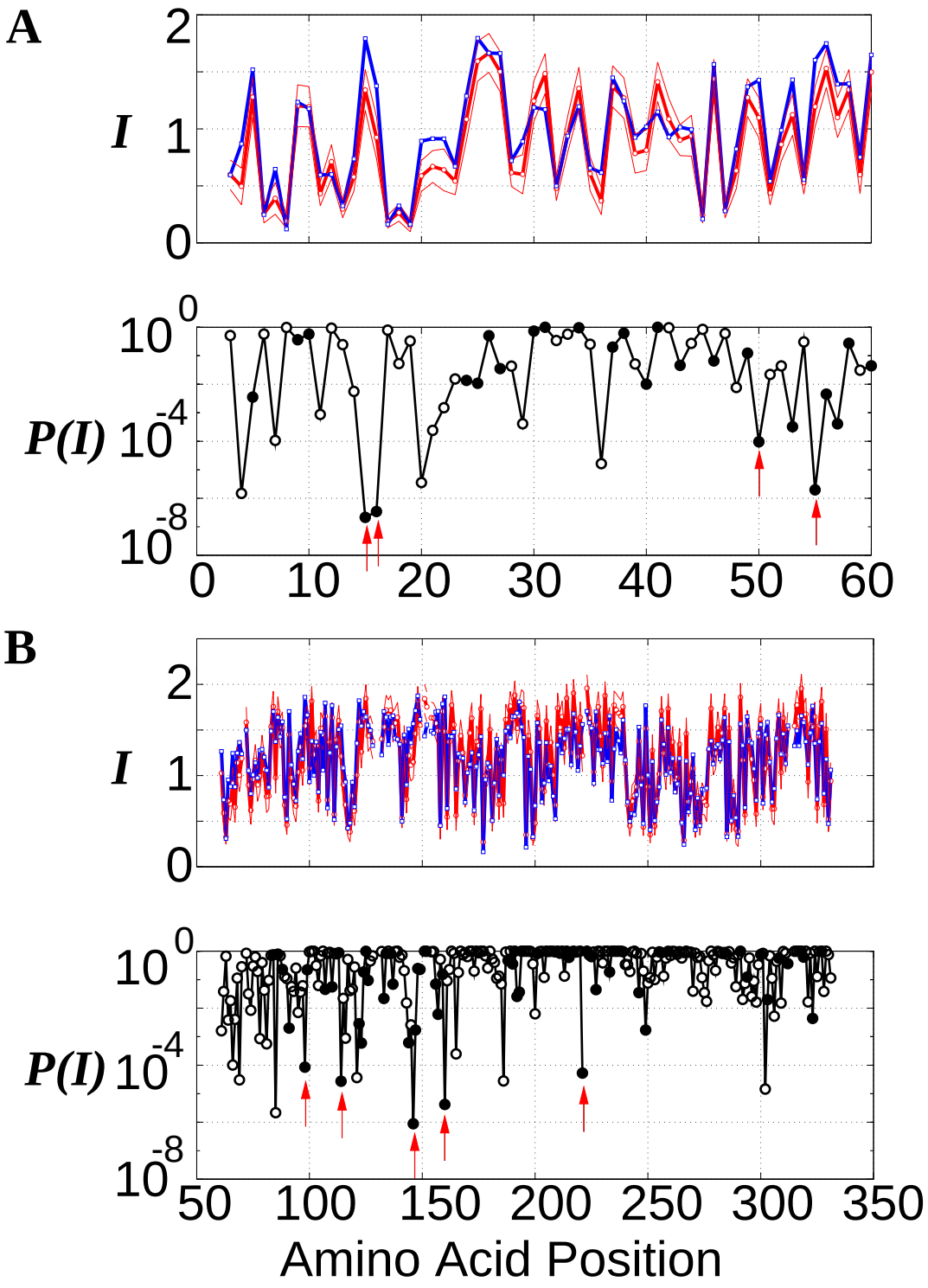
A
I
2
1
0
P(I)
10^0
10^-4
10^-8
0
10
20
30
40
50
60
B
I
2
1
0
P(I)
10^0
10^-4
10^-8
50
100
150
200
250
300
350
Amino Acid Position

Figure 1:

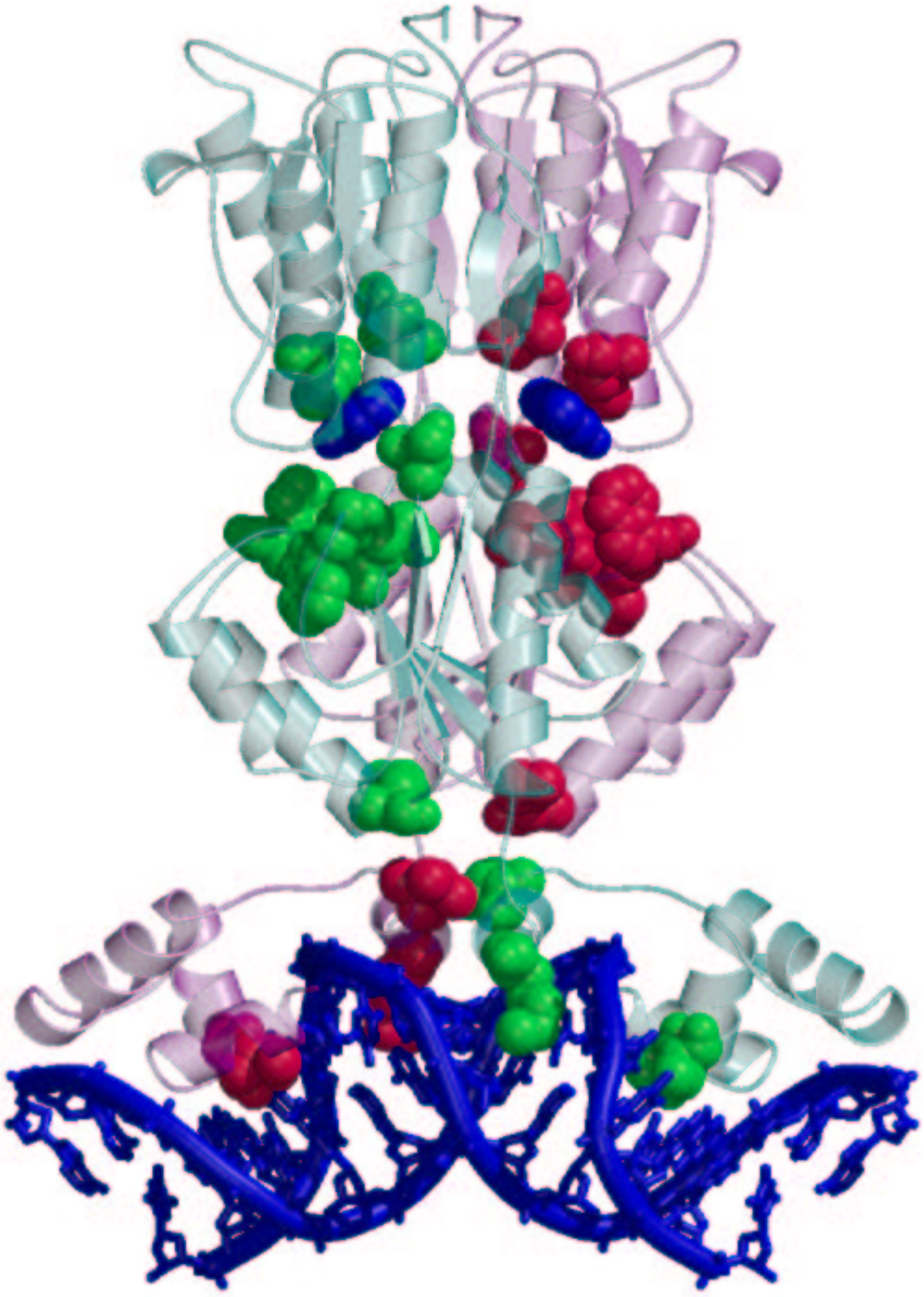

Figure 2:

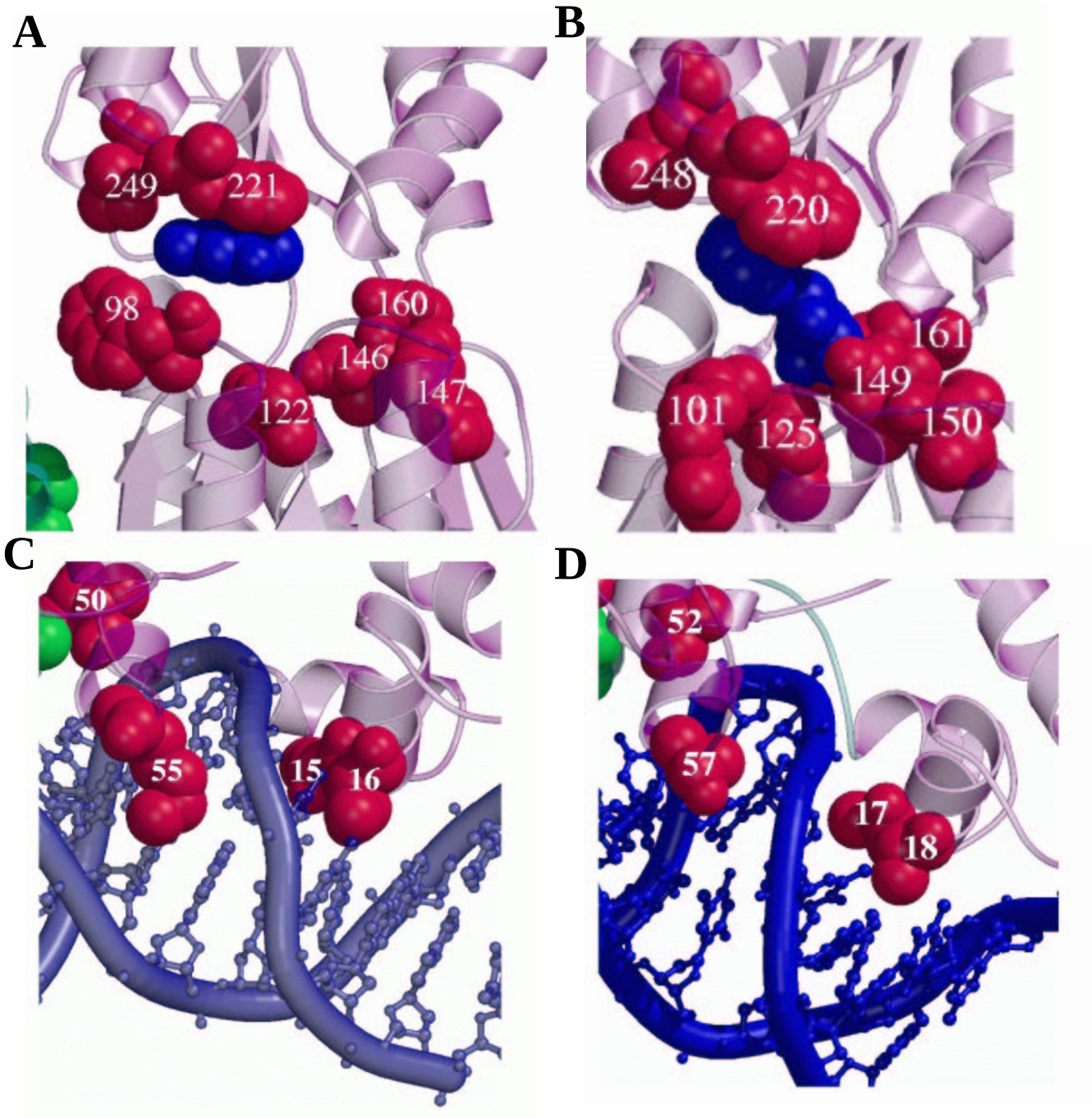
A
249
221
98
160
146
147
122
B
248
220
161
149
101
125
150
C
50
55
15
16
D
52
57
17
18

Figure 3: