i-Patch: Interprotein contact prediction using local network information

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · OCTOBER 2010

Impact Factor: 2.63 · DOI: 10.1002/prot.22792 · Source: PubMed

5 AUTHORS, INCLUDING:

Qiang Luo
Fudan University
28 PUBLICATIONS **94** CITATIONS
SEE PROFILE

Gesine Reinert
University of Oxford
76 PUBLICATIONS **1,014** CITATIONS
SEE PROFILE

Charlotte M Deane
University of Oxford
110 PUBLICATIONS **3,016** CITATIONS
SEE PROFILE

i-Patch: Inter-Protein Contact Prediction using Local Network Information

Rebecca Hamer[1,2,†], Qiang Luo[2,3,†],
Judith P. Armitage[1], Gesine Reinert[1,2], Charlotte M. Deane[1,2,*]

[1] Oxford Centre for Integrative Systems Biology, University of Oxford, Oxford, UK

[2] Department of Statistics, University of Oxford, Oxford, UK

[3] Department of Mathematics and Systems Science, National University of Defense Technology, Changsha, Hunan 410073, China

† These authors contributed equally to this work

* Correspondence to: Charlotte M. Deane, University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, England, OX1 3TG. Email: deane@stats.ox.ac.uk, Telephone: +44(0)1865 281301, Fax: +44(0)1865 272595.

Short title: Inter-protein contact prediction

Key words: protein interactions, correlated mutations, protein complexes, propensities, specificity

Abstract

Biological processes are commonly controlled by precise protein-protein interactions. These connections rely on specific amino acids at the binding interfaces. Here we predict the binding residues of such inter-protein complexes. We have developed a suite of methods, i-Patch, which predict the inter-protein contact sites by considering the two proteins as a network, with residues as nodes and contacts as edges. i-Patch starts with two proteins, A and B, which are assumed to interact, but for which the structure of the complex is not available. However, we assume that for each protein we have a reference structure and a multiple sequence alignment of homologues. i-Patch then uses the propensities of patches of residues

to interact to predict inter-protein contact sites. i-Patch outperforms several other tested algorithms for prediction of inter-protein contact sites. It gives 59% precision with 20% recall on a blind test set of 31 protein pairs. Combining the i-Patch scores with an existing correlated mutation algorithm, McBASC, using a logistic model gave little improvement. Results from a case study on bacterial chemotaxis protein complexes demonstrate that our predictions can identify contact residues, as well as suggesting unknown interfaces in multi-protein complexes.

The i-Patch package is freely available at http://www.stats.ox.ac.uk/research/bioinfo/resources

Introduction

The prediction of binding residues for inter-protein contacts and the identification of specificity-determining residues for these interacting protein pairs would greatly improve our understanding of protein interactions, and have a significant impact in the field of drug discovery. There is very limited experimental data available about which sites in proteins are responsible for the specificity of binding, so testing prediction algorithms on sufficient real examples to obtain significant results is impossible. However, for a given protein complex, the set of inter-protein contact sites will contain a subset of residues which are responsible for the specificity of interaction. Therefore, even if an algorithm accurately predicts only inter-protein contact sites, it will still greatly reduce the set of residues to be tested experimentally for their contribution to interaction specificity. One set of computational techniques for identification of specificity-determining residues are those using correlated mutations. Methods include the Observed Minus Expected Squared (OMES) covariance algorithm [1, 2], which is based on the Chi-squared test, with the null hypothesis that there is no dependence between the two residue distributions in two columns of a multiple sequence alignment (MSA). The McLachlan Based Substitution Correlation (McBASC) method [3, 4] uses the McLachlan substitution matrix to determine the similarity of mutations in two columns of an MSA, and provides a correlation score for every pair of columns. The Mutual Information (MI) algorithm [5], which is based on Shannon's entropy, compares the joint probability distribution of two columns of an MSA to their marginal distributions. High MI values occur if substitutions at two positions in an MSA are correlated. This technique has been used and adapted by many groups [6–11]. H2r [12] is a recently-developed method also based on Shannon's information theory, but which reports residues with high connectivity, *i.e.* those which are involved in many high-scoring pairs. The Statistical

Coupling Analysis (SCA) method [13–15] is based on perturbation of an MSA. It creates sub-alignments which are defined by the most prevalent residue in a column, and asks if this sub-alignment has a changed residue composition compared to the parent alignment from which it was drawn. The Explicit Likelihood of Subset Covariation (ELSC) algorithm [16] is similar to SCA, the difference being in the measurement of deviation of amino acid composition between the sub-alignments and the total alignments. Another technique searches for 'strong motif families' in an MSA, meaning sequences which have a different residue distribution from the rest of the sequences in at least two sites [17]. Other methods attempt to take into account the phylogenetic history of the proteins [8, 18–20].

Most of these algorithms were designed to identify two correlated residues within a single protein sequence (intramolecular correlations) rather than those between two separate proteins (intermolecular correlations). They have generally been evaluated by their ability to predict known intra-protein contact sites in protein structures, but tend to have very low accuracies with high false positive rates [3, 21, 22]. Different algorithms have different sensitivities to background conservation in an MSA [2]. The abilities of OMES, McBASC, MI and SCA to predict intra-protein contact residues were compared in 2004 [2]. Accuracy was defined as the number of correctly predicted contact residues divided by the total number of predictions. Using a test set of 224 protein families, if the top 40 predictions were taken, the accuracies for intra-protein contact prediction for these methods were 13%, 14%, 5% and 5% respectively. A more recent study [22] collated accuracies of intramolecular contacts from several publications, and show that OMES gives an accuracy of 11%, MI between 2% and 18% and SCA between 5% and 7%.

Some correlated mutation prediction methods have been extended to predict intermolecular contacts between protein domains or between proteins, by concatenating two paired sequences [4, 23]. However, the accuracy for prediction of intermolecular pairs is estimated to be approximately ten times lower than that for intramolecular pairs, suggesting that the signals of correlated mutations are weak, and it has been suggested that current methodologies of correlated mutation analysis are not suitable for intermolecular contact prediction [21]. Nevertheless, the MI algorithm has been applied with some success to identify the specific residues in bacterial histidine kinases that determine to which response regulator proteins they bind [23].

It has been suggested that information from correlated mutation analyses should be combined with other sources of information, such as secondary structure and solvent accessibility, to reduce false positive rates [22]. Solvent accessibility of residues has already been used to improve some tools, such as

P2PConPred. This tool uses pair-to-pair substitution probabilities of contact sites for prediction of intra-protein contact residues [24]. It gave a mean accuracy of 13.5%, which increased to 24% when solvent accessibility was used so that only contacts for core residues were predicted.

Some methods have been developed which take into account the fact that multiple residues may mutate simultaneously. Liu *et al.* [25] developed a hidden Markov support vector machine which uses both residue accessible surface area and sequence neighbour information as training features. SCOTCH [26] analyses the physiochemical complementarities between contacting residues at protein complex interfaces. SCOTCH considers a residue with its k-nearest intra-protein neighbours in 3D-space as a patch. Amino acids are classified into 4 categories (hydrophobic, polar, acidic and basic), and three types of complementarity between residues are used (hydrophobic-hydrophobic, polar-polar and oppositely charged). The authors found that complementarity between two interacting proteins may be maintained through matching two patches rather than strictly matching two sites. However, this tool has been used to identify near-native solutions from a set of docking models, rather than to identify contact residues.

We have developed a suite of new methods for inter-protein contact prediction (i-Patch), which detect inter-protein contact sites by considering a protein as a network, with residues as nodes and contacts (both intra- and inter-protein) as edges. Profiles of nodes, edges (pairs of residues) and triangles for inter-domain and inter-protein contact sites in a set of multi-domain proteins and complexes with known structures were calculated. These were shown to be different to those for non-contact sites. Residue, pair and triangle propensity scores for each site can therefore be established by examining the degree to which the residue fits the profiles for contact sites. Three scores, APro, PPro and TPro, were developed using these propensities. The input is two MSAs (one for each of the interacting proteins), along with a known structure for one member of each of the sequence alignments. The structures allow us to identify intra-protein edges (residues within 4.5Å in the structure) and whether residues (nodes) are on the surface of the protein. Scores are assigned to each position in the MSA by using information from all the amino acids present in the MSA column. Each position in the MSA, together with its corresponding column of amino acids, is referred to as a 'site'. APro uses the relative propensity of a single site to be involved in a contact. PPro uses the relative propensity of a pair of sites, one from each protein, to be a contact pair. TPro uses relative contact propensities of triangles of sites (an inter-protein pair plus a third site with an intra-protein edge to one of the other two residues in the reference sequence). All three scores use weights which take into account the intra-protein neighbours of each site, as the structural context

of every residue will affect its propensity to be involved in a contact. All scores are therefore 'patch-based'. We combined the APro, PPro and TPro scores with a score from an existing correlated mutation algorithm, McBASC, using a logistic regression model, but this showed no significant improvement over the use of the TPro score alone. Test on a blind data set show that the TPro score performs best and can achieve a precision 59% at a recall of 20%. Results from a case study on bacterial chemotaxis protein complexes demonstrate that our predictions can identify contact residues, as well as predicting unknown interfaces in multi-protein complexes.

Materials and methods

In all structure data sets, only structures solved by X-ray diffraction, with a resolution of 2.5Å or better, and a chain length of 100 residues or more were used. In addition, the accessible surface area (ASA) of the domains in each protein, or the proteins in each complex, was calculated using JOY [27], and compared to the ASA of the whole docked protein. Any structures with a change in ASA less than 175Å^2 were excluded. This cut-off was established by examining the change in ASA for all the proteins in our domain-domain and protein complex data sets. Very few of these proteins have a change in ASA of less than 175Å^2.

Data for calculating propensities

Domain-domain interactions

SCOP 1.73 [28] was downloaded from http://scop.mrc-lmb.cam.ac.uk/scop-1.73/parse/index.html and all entries with only two annotated domains were selected. Pisces [29] was used to remove proteins with greater than 70% identity, and to exclude entries containing only coordinates for C_α atoms. The final domain-domain data set contains 1150 protein chains (Table 1).

Protein complexes

The advanced search facility from the PDB website [30] was used to obtain a list of PDB entries which contain only protein (no DNA, RNA or hybrids) and have no ligands. Only entries with 2 chains, an oligomeric state of 2 and only 1 model present in the PDB file were selected. Pisces was then used to

remove proteins with greater than 70% identity. The final protein complex data set contains 677 proteins (Table 1).

Amino acid propensities for the domain-domain data set and the protein complex data set were calculated and seen to be highly similar when viewed as probability distributions (based on a Kolmogorov-Smirnov test) (Table 2). Therefore the two data sets were joined together to give the final Propensity data set of 1799 proteins, from which amino acid, pair and triangle propensities were calculated.

Data for fitting models

For both our Fitting and Blind data sets, we use carefully-chosen multi-domain proteins rather than protein complexes, and treat each domain as if it were a separate protein. This enables us to be certain that correct protein-protein pairings are used in the multiple sequence alignments.

Proteins with known structures and annotated domain boundaries were taken from 4 published papers [4, 31–33]. Only proteins for which the annotated domains were clearly defined when the structure was viewed in PyMOL [34] were selected, but any minor errors in domain boundary annotation were corrected after examining the structure. Domain definitions are given in Supplementary file Table S3. Each selected protein was used as a query sequence to BLAST [35, 36] against the nr database [37]. Default settings were used but with the maximum e value set to 1, the low complexity filter turned off and the number of returned hits set to 10000. Hits were only accepted if the hit length was between 75 and 125% of the query length, and if the hit covered the central 50% of the query sequence. For each protein, sequences for all its accepted BLAST hits were obtained from the NCBI protein sequence database. The sequence of the known protein structure was added to this list, then redundancy was removed using CD-HIT [38], with the identity cut-off set to 0.9. The sequence of the known structure was always retained. 31 proteins were selected which had more than 50, and less than 1000, accepted, non-redundant BLAST hits. These 31 sequences were less than 40% identical. This data set is referred to as the Fitting data set. Details are shown in Table 3 and Table S3. The sequence of each known protein structure and its corresponding set of non-redundant BLAST hits were then aligned using MUSCLE [39]. This alignment was submitted to MaxAlign [40] to improve the alignment area, ensuring the query sequence was preserved in the alignment. If sequences were removed by MaxAlign, MUSCLE was then re-run. The known structures of the 31 proteins were used to obtain residues in contact between the domains of each protein, for assessment of algorithm accuracy. This was done by calculating the distance between all atoms in residue

pairs between two domains. If any distance between the atoms of two residues in separate domains is less than 4.5Å these two residues are said to be in contact [41] and are used as the 'True Positive' set of contact residues.

Data for blind testing

PDB data

The advanced search facility from the PDB website was used to obtain a list of PDB entries which contain only protein (no DNA, RNA or hybrids) and have no ligands. This is referred to as the Gold Standard PDB list and is used to filter data taken from SCOP [28] and CATH [42] as described below:

SCOP and CATH data

SCOP 1.73 and 1.75 were downloaded from http://scop.mrc-lmb.cam.ac.uk. All entries present in SCOP 1.75 but not in SCOP 1.73 were extracted to avoid overlap with data used in the propensity data set. Those with two continuous domains and which correspond to a PDB entry in the Gold Standard PDB list were selected. Any proteins annotated as T-cell receptor (TCR), human leukocyte antigen (HLA) or antibody (Ig) were rejected as these may have relatively low sequence identity but are structurally very similar. This left 32 proteins which were examined using PyMOL. Any with SCOP domain definitions that were unclear were removed, leaving 21 proteins.

CathDomall version 3.2.0 was downloaded from http://www.cathdb.info/wiki/data:index. 116 entries were selected which have only one chain, and two continuous domains, but are not present in SCOP 1.75 and correspond to a PDB entry in the Gold Standard PDB list. Any CATH entries annotated as T-cell receptor (TCR), human leukocyte antigen (HLA) or antibody (Ig) were removed.

Final blind data set

The filtered SCOP and CATH entries were joined together. CD-HIT-2D was then used to select proteins with less than 70% identity to proteins in the propensity data sets and the 31 proteins in the Fitting data set. Pisces was run on the remaining proteins using an identity cut-off of 30%. All remaining proteins from SCOP and a random selection of the proteins remaining from CATH were then used as BLAST queries against the nr database, and the results filtered as described previously. Any proteins with fewer

than 100 accepted BLAST hits were rejected. CD-HIT with a 90% identity cut-off was then run on each set of BLAST hits. A final set of proteins was selected which had more than 100 and fewer than 1000 accepted, non-redundant BLAST hits. This data set is referred to as the Blind data set and contains 31 proteins. Details are shown in Table 3 and Table S4. The sequence of each known protein structure and its corresponding set of BLAST hits were aligned using MUSCLE and MaxAlign, and the PDB structures of the 31 proteins were used to define residues in contact between the respective domains, as described for the Fitting data set.

Case study data

The five-domain bacterial chemotaxis protein CheA is known to interact with CheY via its first domain, P1, and with CheW via its fifth domain, P5 [43–45]. It is generally assumed that proteins encoded by genes within the same operon are likely to interact. This assumption was used to obtain a large set of CheAP1-CheY and CheAP5-CheW protein pairs as follows: genes encoding the core chemotaxis proteins (CheA, CheB, CheR, CheW and CheY) from 474 non-redundant complete bacterial genomes were clustered into putative operons, as previously described [46]. Each CheA-CheY and CheA-CheW pair encoded in the same operon was extracted. The P1 and P5 domains from the CheA proteins were established by their similarity to the respective domains in *E. coli* CheA.

For CheAP5-CheW predictions, the crystal structure of the CheAP4-CheAP5-CheW complex from *Thermotoga maritima* (PDB identifier 2CH4, [45]) was used to define exposed/buried residues for each protein, as well as to identify residues in contact between the respective domains (true positive contact sites). For CheAP1-CheY6 predictions, the structures of CheAP1 (PDB identifier 1TQG, [47]) and CheY (PDB identifier 1TMY, [48]), both from *Thermotoga maritima* were used in the alignment to give exposed/buried information for each site.

The sequences of the known protein structures and their corresponding homologues of each protein/domain (CheAP1, CheAP5, CheW and CheY) were aligned using MUSCLE and MaxAlign. The domain pairings between sequences were then assigned based on the operon information described above, and paired sequences from the alignments were concatenated. Only one-to-one pairings were used, *i.e.*, if a gene encoding CheA was present in an operon with several genes encoding CheYs, only one of these pairings was used. Redundancy was then removed using CD-HIT with an identity cut-off of 90%. The multiple sequence alignments used by i-Patch in this paper for the CheAP5-CheW and CheAP1-CheY6

in Texts S2 and S3, respectively.

Exposed/buried information

For each protein in our data sets, we calculate the solvent accessibility for the residues on each domain by considering each domain as a separate molecule. A surface-exposed residue on our reference structure must have at least 7% of its side chain accessible to a 1.4Å radius probe (as defined by the program PSA, run as part of JOY [27]). This information about a residue is then annotated to the entire MSA column to which it corresponds. Thus each site can be described as either exposed or buried. We define a 'site' as a position in the MSA, together with all the residues in its corresponding MSA column. If the reference sequence has a gap at any site, this site is assumed to be exposed.

Contact residues at the interface between the two domains must be surface exposed. Any site which is buried can therefore be given a score of 0 and excluded. For a fair comparison, we run the standard correlated mutation algorithms MI, McBASC, ELSC, OMES and SCA, now with exposed/buried (EB) information included. We call these EB scores (EBMI, EBMcBASC, EBELSC, EBOMES, EBSCA).

Definition of contacts

An inter-protein contact site is defined as a surface-exposed residue which is less than 4.5Å away from another surface-exposed residue on a different protein (taking all atoms into account).

An inter-protein contact pair is defined as two surface-exposed residues on different proteins, which are less than 4.5Å away from each other. A non-contact pair is defined as two surface-exposed residues on different proteins which are at least 4.5Å apart. Background pairs consist of all the possible surface-exposed residue pairs.

We define triangles as three surface-exposed residues from two different proteins, where two of the residues are present on the same protein, less than 4.5Å apart from each other, and a third residue is on the surface of the other protein. A contact triangle is defined when the distance between each pair of sites in the triangle is less than 4.5Å; otherwise, if the distance between some pair of sites in the triangle is 4.5Å or greater, it is a non-contact triangle.

The intra-protein neighbours of a residue are defined as those surface-exposed residues in the same protein, which are less than 4.5Å away from the residue of interest.

Calculation of amino acid propensities

For each of the 20 standard amino acids, a, let f_a^{con}, f_a^{non}, and f_a^{all} denote the frequencies of amino acids on the surface of proteins, at contact sites, non-contact sites, and all sites, respectively, in our Propensity data set. The amino acid propensities for contact sites, p_a^{con}, and non-contact sites, p_a^{non}, on the surface of proteins are calculated as follows:

$$p_a^{con} = \frac{f_a^{con}/f_a^{all}}{\sum_b f_b^{con}/\sum_b f_b^{all}} \quad and \quad p_a^{non} = \frac{f_a^{non}/f_a^{all}}{\sum_b f_b^{non}/\sum_b f_b^{all}}.$$

The relative propensity of an amino acid, a, to be a contact site on the protein surface is calculated as

$$p_a^{con}/p_a^{non}. \tag{1}$$

Figure 1A shows the propensities for each amino acid to be in contact sites, non-contact sites and their relative propensities. For example, C, I, L, M, F, W, Y and V have considerably higher propensities to be contact sites than to be non-contact sites, when surface-exposed.

Pair and triangle propensities

If the set $\Sigma_{aminoacid}$ of all 20 amino acids is taken into account, there are 210 different pairs of residues, and 1540 different triangles. To establish reliable propensities for those would require a very large data set. Therefore, we group the 20 amino acids into a set, $\Sigma_{category}$, of 7 categories according to their physicochemical properties: Small (S,G,A,P), Hydrophobic (V,M,I,L,C), Negatively charged (D, E), Aromatic (F,Y,W), Polar (Q,T,N), Favoured Positively-charged (R,H) and Disfavoured Positively-charged (K). These categories are abbreviated to S, H, N, A, P, fP and dfP. As Lysine (K) is found to be rare at protein/domain interfaces (propensity 0.66), while Arginine (R) and Histidine (H) are much more common (propensities of 1.05 and 1.11, respectively), a division into Favoured and Disfavoured categories is introduced. Each amino acid, a, is assigned to a category, C.

Pair propensities

Instead of pairs of amino acids (a_1, a_2), pairs of categories of amino acids, (C_1, C_2), from the 7 categories described above (S, H, N, A, P, fP and dfP) are considered. There are 28 possible types of pairs,

$\Sigma_{pair} = \{SS, SH, SN, SA, SP, SfP, SdfP, HH, HN, \cdots, dfPdfP\}$. The frequency f_P^{con} of the pair $P = (C_1, C_2) \in \Sigma_{pair}$ occurring as a contact pair can be calculated by counting the number of contact pairs (a_1, a_2), where a_1 belongs to the category C_1 and a_2 belongs to the category C_2. Similarly, the frequencies for non-contact pairs, and background pairs are denoted as f_P^{non} and f_P^{all}, respectively. For any pair $P \in \Sigma_{pair}$, the propensity of contact pairs, p_P^{con}, and the propensity of non-contact pairs, p_P^{non}, are calculated in an analogous way to that of amino acid propensities:

$$p_P^{con} = \frac{f_P^{con}/f_P^{all}}{\sum_{Q\in\Sigma_{pair}} f_Q^{con}/\sum_{Q\in\Sigma_{pair}} f_Q^{all}} \quad and \quad p_P^{non} = \frac{f_P^{non}/f_P^{all}}{\sum_{Q\in\Sigma_{pair}} f_Q^{non}/\sum_{Q\in\Sigma_{pair}} f_Q^{all}}$$

and the relative pair propensity for contact pair P is

$$p_P^{con}/p_P^{non}. \tag{2}$$

Note that for the pair propensity, the ordering of the amino acids in a contact pair does not matter.

Triangle propensities

The triangle T is defined using the residue categories C_1, C_2, and C_3 as described above, and the set of possible triangles is denoted by $\Sigma_{triangle}$ = { SHN, SHA, $\cdots$, PfPdfP, SHH, SSH, $\cdots$, fPfPdfP, SSS, HHH, $\cdots$, dfPdfPdfP}. This set contains 84 different triangle types. Propensities p_T^{con} and p_T^{non} are calculated in an analogous way to that of amino acid propensities, by counting the frequency of triangle type T to be a contact triangle (f_T^{con}) and the frequency for it to be a non-contact triangle (f_T^{non}). For triangle type $T = (C_1, C_2, C_3)$, the relative triangle propensity is:

$$\frac{p_T^{con}}{p_T^{non}} = \frac{f_T^{con}/f_T^{non}}{\sum_{R\in\Sigma_{triangle}} f_R^{con}/\sum_{R\in\Sigma_{triangle}} f_R^{non}}. \tag{3}$$

Again, the ordering of the three amino acids within a triangle does not affect the relative triangle propensity.

Calculation of scores

Intra-protein weights (w_{intra})

The set of intra-protein neighbours of a residue, a, is denoted by

$$N(a) = \{b \text{ residues in the same protein as } a : d_{a,b} < 4.5\mathring{A}, b \text{ is a surface} - \text{exposed residue}\},$$

where $d_{a,b}$ is the distance between residues a and b in space.

The intra-protein neighbours of a residue category, C, can also be defined as follows; Let $f_{C_1}^{con}$ denote the frequency of residues in category C_1 occurring as contact sites in our propensity data set. Let $f_{C_2 \in N(C_1)}$ denote the frequency of residues in category C_2 to be intra-protein neighbours of residues in category C_1; $f_{C_1}^{non}$, $f_{C_2 \in N(C_1)}^{non}$, $f_{C_1}^{all}$; $f_{C_2 \in N(C_1)}^{all}$ are the corresponding concepts for non-contact residue categories and all surface residue categories, respectively. The propensities of residues in category C_2 to be intra-protein neighbours of residues in category C_1 for contact sites and non-contact sites are defined as:

$$w_{intra}^{con}(C_2 \mid C_1) = \frac{f_{C_2 \in N(C_1)}^{con} / f_{C_2 \in N(C_1)}^{all}}{f_{C_1}^{con} / f_{C_1}^{all}} \text{ and } w_{intra}^{non}(C_2 \mid C_1) = \frac{f_{C_2 \in N(C_1)}^{non} / f_{C_2 \in N(C_1)}^{all}}{f_{C_1}^{non} / f_{C_1}^{all}}.$$

The relative intra-protein weight is defined as

$$w_{intra}(C_2 \mid C_1) = \frac{w_{intra}^{con}(C_2 \mid C_1)}{w_{intra}^{non}(C_2 \mid C_1)}. \tag{4}$$

A heat-map showing the relative intra-protein weights for each residue category is shown in Figure 3A.

Weights for residue pairs (w_{pair})

Let C_1, C_2, and C_3 be residue categories in $\Sigma_{category}$. The set of intra-protein neighbours of the pair $P = (C_1, C_2)$ is defined as the set of all categories which are neighbours of C_1, C_2 or both:

$$N(P) = N(C_1) \cup N(C_2).$$

We denote by $f_{C_3 \in N(P)}^{con}$ the frequency of residues in category C_3 occurring as neighbours of the contact pair P, and by f_P^{con} the frequency of the pair type P occurring as a contact pair. The frequencies

$f^{non}_{C_3 \in N(P)}$, f^{non}_P, $f^{all}_{C_3 \in N(P)}$, and f^{all}_P are the corresponding frequencies when P occurs as a non-contact pair or a surface pair.

For a pair type $P = (C_1, C_2)$, the propensity of a residue in category C_3 to be an intra-domain, surface-exposed neighbour of *either* C_1 *or* C_2 is calculated as:

$$w^{con}_{pair}(C_3 \mid P) = \frac{f^{con}_{C_3 \in N(P)} / f^{all}_{C_3 \in N(P)}}{f^{con}_P / f^{all}_P} \text{ and } w^{non}_{pair}(C_3 \mid P) = \frac{f^{non}_{C_3 \in N(P)} / f^{all}_{C_3 \in N(P)}}{f^{non}_P / f^{all}_P}.$$

The relative pair weight is then defined as:

$$w_{pair}(C_3 \mid P) = \frac{w^{con}_{pair}(C_3 \mid P)}{w^{non}_{pair}(C_3 \mid P)}. \tag{5}$$

A heat-map showing the relative pair weights for each type of residue pair is shown in Figure 3B.

Propensity Scores

We start with two proteins, A and B, which are assumed to interact, but for which a complex structure is not available. We assume that we have a reference structure and a multiple sequence alignment of homologues for each protein. We recommend that the sequences in each MSA share no more than 90% sequence identity.

The sequences in the two MSAs are concatenated, based on knowledge about which pairs of proteins interact. Following concatenation, all columns which contain 50% or more gaps are removed.

We define a 'site' as a position in the MSA, together with all the residues in its corresponding MSA column. The reference structures are used to assign exposed/buried status to each site, i. If the reference sequence has a gap at any site in the MSA, this site is assumed to be exposed. We introduce three propensity-based scores (APro, PPro and TPro) using the amino acid, pair and triangle propensities from equations (1), (2) and (3) (see also Figures 4, S8, and S9). Propensity scores are calculated for each surface-exposed site, and all buried sites are assigned scores of 0. Let a_{ij} be the residue in sequence j at site i, and let d_{ii_t} denote the distance between two sites, i and i_t, on the same protein. We define the *patch* for site i (the i-patch) as

$$\Pi(i) = \{i_t \text{ is surface exposed on the same protein: } d_{i,i_t} < 4.5\text{Å}\} \tag{6}$$

and $|\Pi(i)|$ is the number of sites in the set $\Pi(i)$. The distance between site i and itself is zero, so $i \in \Pi(i)$, and the weight of the propensity score for site i itself is always 1.

To take gaps in the MSA into account, we define $G(i) = \{j : a_{ji} = '-', j = 1, 2, \cdots, M\}$ as the index set of the sequences which have gaps at site i in the MSA. The index set of gap free sequences at site i is denoted as $G(i)^c = \{j : j \notin G(i), j = 1, 2, \cdots, M\}$.

For all scores, the intra-protein weight, w^{intra}, for the pair of amino acids a_{ji_t} and a_{ji} is calculated using the corresponding residue categories, C_{ji_t} and C_{ji}, as in equation (4). The average intra-protein weight between sites i and i_t is calculated as

$$w_{ii_t}^{\text{intra}} = \frac{1}{M - |G(i_t) \cup G(i)|} \sum_{j \in G(i_t)^c \cap G(i)^c} w^{\text{intra}}(C_{ji_t}|C_{ji}). \tag{7}$$

APro Score

Considering the patch information $\Pi(i)$ on protein A alone, for each surface-exposed site i on protein A we define the patch-based amino acid propensity $S_{i_t}^A$ for each site $i_t \in \Pi(i)$ as

$$S_{i_t}^A = \frac{1}{M - |G(i_t)|} \sum_{j \in G(i_t)^c} (p_{a_{ji_t}}^{con} / p_{a_{ji_t}}^{non}). \tag{8}$$

We then take the weighted average to form the amino acid propensity score, APro:

$$S_i^{\text{APro}} = \frac{1}{|\Pi(i)|} \sum_{i_t \in \Pi(i)} w_{ii_t}^{\text{intra}} S_{i_t}^A. \tag{9}$$

PPro score

The pair propensity of the two amino acids, (a_{ji_t}, a_{jk}), which are on different proteins, is given by the pair propensity of the corresponding two residue categories, C_{ji_t} and C_{jk}, as in equation (2).

The average pair propensity $S_{i_t}^{Pair}$ among all possible pairs between site $i_t \in \Pi(i)$ on protein A and the surface exposed sites on protein B is calculated as

$$S_{i_t}^{Pair} = \frac{1}{|\{k \text{ exposed on B}\}|} \sum_{k \text{ exposed on B}} \frac{1}{M - |G(i_t)|} \sum_{j=1}^{M-|G(i_t)|} (p_{(C_{ji_t}, C_{jk})}^{con} / p_{(C_{ji_t}, C_{jk})}^{non}).$$

For each surface-exposed site i on protein A, the pair propensity (PPro) score is then calculated as

$$S_i^{PPro} = \frac{1}{|\Pi(i)|} \sum_{i_t \in \Pi(i)} w_{ii_t}^{\text{intra}} S_{i_t}^{Pair}.$$

TPro score

The triangle propensity of the three amino acids a_{ji_t}, a_{jk} and a_{jl} is given by the triangle propensity of the corresponding three residue categories C_{ji_t}, C_{jk} and C_{jl}, as in equation (3). The relative pair weight is given in equation (5).

The average triangle propensity, $S_{i_t}^T$, among all possible triangles between site $i_t \in \Pi(i)$ on protein A, the k surface exposed sites on protein B, and a residue, l, which is a structural neighbour of either i_t or k, is calculated as

$$\begin{aligned} S_{i_t}^{Triangle} &= \frac{1}{|\{k \text{ exposed on B}\}|} \sum_{k \text{ exposed on B}} \frac{1}{|\Pi(i_t) \bigcup \Pi(k)|} \cdot \\ &\sum_{l \in \Pi(i_t) \bigcup \Pi(k)} \frac{1}{M - |G(i_t)|} \sum_{j=1}^{M-|G(i_t)|} w^{pair}(C_{jl}|C_{ji_t}, C_{jk})(p^{con}_{(C_{ji_t},C_{jk},C_{jl})}/p^{non}_{(C_{ji_t},C_{jk},C_{jl})}). \end{aligned}$$

For each surface-exposed site i on protein A, the triangle propensity (TPro) score is then calculated as:

$$S_i^{TPro} = \frac{1}{\Pi(i)} \sum_{i_t \in \Pi(i)} w_{ii_t}^{\text{intra}} S_{i_t}^{Triangle}.$$

Logistic regression model

If q_i is the probability of site i to be a contact site, the APro, PPro, TPro and EBMcBASC scores can be combined in the logistic regression model

$$\begin{aligned} \text{logit}(q_i) &= \ln(\frac{q_i}{1-q_i}) \\ &= \beta_0 + \beta_1 \text{APro} + \beta_2 \text{PPro} + \beta_3 \text{TPro} + \beta_4 \text{EBMcBASC}, \end{aligned}$$

where $\beta_i, i = 0, 1, 2, 3, 4$ are the model coefficients. The model coefficients are then estimated by maximum likelihood using the Fitting data set, assuming that the data come from independent Bernoulli trials. This

assumption is an approximation, hence statistical inference is limited, but we can use the fit to calculate the combination score, Comb, as

$$\text{Comb} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1\text{APro} + \hat{\beta}_2\text{PPro} + \hat{\beta}_3\text{TPro} + \hat{\beta}_4\text{EBMcBASC})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\text{APro} + \hat{\beta}_2\text{PPro} + \hat{\beta}_3\text{TPro} + \hat{\beta}_4\text{EBMcBASC})},$$

where $\hat{\beta}_i, i = 0, 1, 2, 3, 4$ are the estimated model coefficients.

PROC curves

Since the contact sites form only 16.17% of all residues in the Fitting data set, the PROC (Precision Recall Operating Characteristic) curve [49] will be more informative than a traditional ROC (Receiver Operating Characteristic) curve [50], especially when the score cut-off is high. The percentiles of the scores are used as the cut-offs to calculate the precision and recall. Precision and recall are defined as follows:

$$\text{precision} = \frac{TP}{TP + FP} \; \textit{and} \; \text{recall} = \frac{TP}{TP + FN}$$

Results and Discussion

Data

Four data sets for parameterising and testing the method were collated. First, a set of 1150 two-domain protein chains, and second a set of 677 protein complexes were used to calculate the propensities of amino acids to occur as contact residues, contact pairs and contact triangles (Table 1). These two data sets were then combined, as the propensities obtained from the two data sets were highly similar (based on a Kolmogorov-Smirnov test [51]) (Table 2). Thus over 100 000 contact residues were used in the calculation of our amino acid, pair and triangle propensities. The third set, of 31 proteins, is our Fitting data set and is a set of structurally solved, well-annotated, multi-domain proteins (Table 3). For these 31 proteins, an MSA was generated for each by using BLAST [35,36] to obtain homologues from NCBIs nr database [37]. The fourth set, also of 31 proteins, is our Blind data set, and was constructed from SCOP [28] and CATH [42] entries containing two continuous domains (Table 3). Again, for all 31 proteins, MSAs were generated as for the Fitting data set.

The structure and domain definitions of each protein in our Fitting and Blind data sets were checked

by hand to ensure the domains were separate structural units. The Blind data set was used to ensure we tested our combined score fairly, as the model parameters for this score were estimated using the Fitting data set, but all algorithms were tested on both data sets for comparison.

Exposed/buried information improves prediction of existing algorithms

The ability of five established correlated mutation algorithms (MI, McBASC, ELSC, OMES and SCA) to predict inter-protein contact sites was tested on the Fitting data set. This consists of 31 proteins with at least two domains, containing 1690 inter-domain contact sites (Table 3 and Table S3). McBASC achieves the best result, with 24.5% precision at at recall of 20% (Figure 2A). Precision measures the number of correct contact site predictions out of the total number of predictions made, and recall measures the number of correct contact site predictions out of the total number of real contact sites.

All the algorithms were improved by the inclusion of exposed/buried (EB) information, where no score is given to buried residues. EBMcBASC still gave the best results. MI is least affected by EB information, with the precision only rising from 20% to 25% at a recall of 20%, whereas the precision for McBASC rises from 24.5% to 36% at 20% recall (Figure 2B). The areas under the ROC curves (supplementary material) were also compared, confirming that the EB information improves all the prediction algorithms. The same significant improvement when using EB information can also be seen in the results on the Blind data set (supplementary material). We add EB information to all subsequent scores, and only assign scores to residues which are surface-exposed.

Predicting contacts using patches of residues

i-Patch is used to predict binding sites between two proteins which are assumed to interact, for which a reference structure and a multiple sequence alignment of homologues are available for each protein. The sequences in the two MSAs are then concatenated, based on knowledge about which pairs of proteins interact, and the reference structures are used to assign exposed/buried status to each site, where a 'site' is a position in an MSA, together with all the residues in that MSA column.

A very simple score can be built from the propensities of each of the 20 amino acids to be a contact residue, and a score for each site can be obtained by combining the propensities from the whole MSA column. We could use these propensities alone to score the odds of a site to be involved in an inter-protein contact. However, if we examine the amino acid composition of the top 40 predictions from such a score

(described in supplementary material), it is clear that it is very different from the amino acid composition of real contact residues in our Fitting data set (Figure 1). Hydrophobic and aromatic residues have a far higher propensity to be involved in inter-protein contacts than other residue types. These amino acids drown out other residues, despite the fact that all types of amino acids can be involved in contacts, depending on their structural context.

In reality, interaction between two proteins is usually mediated through patches of residues on the exposed surfaces of proteins. i-Patch therefore considers a protein structure as a network where nodes are residues and edges are contacts. Two types of contacts are considered; intra-protein and inter-protein. We introduce the intra-protein weight (w_{intra}) to integrate the information from the intra-domain neighbours in 3D-space of each site, and give the first i-Patch score, APro. The intra-protein weight is established by calculating the propensities of amino acid types to be neighbours of a contact site residue. The intra-protein neighbours of a site are defined as residues less than 4.5Å away from the residue of interest in the reference structure. In this case, 7 amino acid categories were used, as a very large data set would be needed to establish reliable propensities for all 20 amino acids. For example, if a contact residue is aromatic, there is a high propensity that an intra-neighbouring residue will be hydrophobic. Conversely, if a contact residue is hydrophobic, there is a low propensity that an intra-neighbouring residue will be negatively charged (Figure 3A and Methods). As with the scores for individual sites, all residues in the MSA columns corresponding to neigbouring residues in the reference structure are used to calculate the overall intra-protein weight. The amino acid composition of the top 40 predictions given by this weighted patch-based score approximates the composition of the real contact sites in the Fitting data set far better (Figure 1) than the non patch-based score.

The second and third i-Patch scores, PPro and TPro, use the propensities of 28 types of inter-protein contact pairs and 84 types of inter-protein contact triangles, respectively, along with w_{intra}, to give a score for each site. In addition to w_{intra}, TPro also uses another weight, w_{pair}. This takes into account the residue types surrounding a pair of amino acids involved in an inter-protein contact (see Methods) (Figure 3B). For example, if a pair of residues (one from each protein) are in contact, and both residues are small, there is a high propensity that a neighbouring residue in one or other of the proteins is hydrophobic. The three scores are summarized in Figures 4, S8 and S9.

Comparison of i-Patch scores with existing algorithms

We compared the ability of our patch-based propensity scores to predict contact sites to that of five established correlated mutation algorithms. Our scores perform far better on both the Fitting and Blind data sets (Figure 5). On the Fitting data set, at 20% recall, the APro, PPro, and TPro can achieve 54%, 53%, and 59% precision, respectively, compared to a precision of 35% for EBMcBASC (Figure 5a). On the Blind test data set, the results are similar (Figure 5b), with the APro, PPro and TPro scores having precision of 51%, 49%, and 52%, respectively, at 20% recall, while EBMcBASC has a precision of 28%.

We also combined our propensity scores, along with the EBMcBASC score, using a logistic regression model, to see if this gave any further improvement. We chose EBMcBASC as it performed best out of the five correlated mutation algorithms on the Fitting data set. The model coefficients were estimated using the Fitting data set (Table 4) and the resulting score is called the Combined score (Comb). More details about the logistic model can be found in the supplementary material. Figures 5a and 5b show the performances of TPro and Comb on the Fitting and Blind data sets respectively. When using the Fitting data set, slightly better precision can be gained by this combination compared with the propensity scores alone, but on the Blind test data set, the advantage of the combined score is not as clear. We have created a package for inter-protein contact prediction, i-Patch, which outputs all four scores, but recommend the use of the TPro score as it has similar performance to the combined score, but does not rely on a fitted model.

Finally, we compared iPatch scores to those from a newly-published modification of the MI score, aMIc [11]. These results are shown in supplementary figures S12-S14. The performance of aMIc on the Blind data set is comparable to that of the other algorithms, both with and without EB information. iPatch scores outperform aMIc in both cases.

Case studies: i-Patch predictions on bacterial chemotaxis protein complexes

Many bacteria are chemotactic, meaning that they can sense their chemical environment and move towards more favourable conditions. The chemotaxis system in *Escherichia coli* has been extensively studied and is well understood [52]. Briefly, membrane-spanning methyl-accepting chemotaxis receptors (MCPs) sense any reduction in chemoattractant or increase in chemorepellent in the environment. The MCPs are linked via CheW to a dimeric histidine protein kinase (HPK), CheA. On activation, the

monomers of CheA transautophosphorylate and the phosphate group can then be transferred to the response regulator CheY. Phosphorylated CheY is able to diffuse and bind to the flagellar motor, resulting in a change in motor rotation direction. This causes a switch from smooth swimming to tumbling, allowing the bacteria to change direction. The chemotaxis pathway is conserved across most bacterial species, and many species have more than one, apparently homologous, pathway. Chemotaxis is therefore an ideal system on which to predict protein:protein interactions, and one for which the identification of specificity-determining residues would be extremely useful.

As there is no clear correlation between the total length of two proteins/domains and the number of contact sites between them (supplementary material), we select the top 40 predictions from our scores for both our case studies.

i-Patch predictions on CheA P5 domain - CheW

The classical CheA protein has five domains designated P1 to P5 [43,53]. Homologues of this protein are easily identifiable, as this domain structure is different to other HPKs in bacteria. P1 is the histidine-containing phosphotransfer (HPt) domain, P2 contains the binding site for CheY and CheB, P3 is the dimerization domain, P4 is the kinase domain which phosphorylates a conserved histidine in P1, and P5 binds to CheW and the receptors. No structure of the entire CheA domain has yet been solved, but a structure of the CheA P4 and P5 domains from *Thermotoga maritima* in complex with CheW is available (PDB identifier 2CH4; [45]). We used this structure to identify which residues are in contact between the CheA P5 domain and CheW, and compare these to the predictions given by i-Patch (Figure 6). The top 40 TPro scores predict 13 contact sites correctly, out of a total of 36 real contact sites across the two proteins. 27 predictions do not correlate with real contact sites between P5 and CheW. However, if the complex of CheA P5-CheW is superimposed onto the complex of CheA P3, P4 and P5 domains from *Thermotoga maritima* (PDB identifier 1B3Q; [43]), it can be seen that 6 false positive predictions appear to be at the interface between the CheA P5 and P3 domains, and 10 at the region that is proposed to bind to the MCPs [45] (Figure 7). Some of the other 'false positive' predictions may therefore give clues as to where the P1 and P2 domains of CheA fit into the complex.

i-Patch predictions on CheA P1 domain - CheY

Upon activation, the P1 domain of CheA transfers a phosphate group to CheY. We used i-Patch to predict interactions between the CheA P1 domain and the CheY protein, using reference structures from *Thermotoga maritima*. Prediction results are shown on the sequence of the *T. maritima* proteins in Figure 8, and Figure 9 shows the TPro predictions mapped onto the *T. maritima* reference structures used in the MSA. These predictions can be verified when the structure of CheAP1-CheY6 from *Rhodobacter sphaeroides* is released by the PDB (PDB identifier 3KYI, currently awaiting processing).

It can be seen that one face of the CheA P1 domain is not predicted to have any binding sites, but that the region near the N-terminus (patch 1) and a large patch made up from residues from two of the helices (patch 2) are predicted to be involved in binding. It is possible that one of these regions binds to CheY and the other to another CheA domain.

There are three patches of predicted contact residues on CheY. Patch 1 (residues 100-104) contains Phe101, a residue equivalent to Tyr106 in *E. coli*, which is known to be essential for signalling ability of CheY [54]. This patch is involved in binding to the P2 domain of CheA in the complex of CheAP2-CheY from *Thermotoga maritima* (1U0S, [55]). Patch 2 (residues 57, 58, 61 and 62) corresponds to loop 5 (residues 55-61) of CheY in *Thermotoga maritima*. This loop contributes a ligand to the metal-binding site of CheY [48] and is highly conserved among CheY homogues. It has been suggested that it is a 'universal recognition element' across the CheY superfamily for recognition by kinases [56]. Patch 3 consists of residues 11-14, 16, 17 and 20. It is likely that either this patch, or patch two,is involved in binding to the CheA P1 domain. Residue Asn116 is not in a patch therefore is likely to be a false positive result.

As a comparison to predictions made by other algorithms [11, 23], i-Patch predictions for a bacterial histidine kinase - response regulator complex are shown in supplementary figure S10.

Conclusions

We have developed a suite of novel methods, i-Patch, to predict inter-protein contact sites by considering a protein as a network, with residues as nodes and contacts as edges. Using a large data set of domain-domain and protein-protein interactions, the propensities of amino acids to be involved in inter-protein contacts, inter-protein contact pairs and inter-protein contact triangles were established. i-Patch uses

these propensities, together with the structural context of each node, to score the odds of a site to be involved in an inter-protein contact. i-Patch takes as input a multiple sequence alignment, which includes two reference structures for the pair of interacting proteins. The reference structures are used to give intra-protein edges and exposed/buried information for the nodes of each protein. It outputs four scores, APro, PPro, TPro and Comb. On our test data sets, all four i-Patch scores significantly outperform five established correlated mutation detection algorithms for detecting inter-protein contact sites. TPro gives a precision of 59% at 20% recall, compared to a precision of 35% at 20% recall for the best-performing correlated mutation algorithm, EBMcBASC.

The i-Patch scores were applied to two case studies to predict the inter-protein contact sites between the chemotaxis proteins CheAP5-CheW and CheAP1-CheY. Comparing our predictions with the contact sites given by the known structure of the CheAP5-CheW protein complex demonstrates that our predictions are successful at identifying the contact sites. Some false positive predictions made by i-Patch may in fact highlight residues involved in other interfaces in the CheA-CheW-receptor complex and the CheA-CheY complex.

Acknowledgments

The authors would like to acknowledge Dr George Wadhams, Dr Sonja Pawelczyk, Dr Kathryn Scott and Dr Steven Porter for helpful discussion about this work.

References

[1] Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins: Structure, Function, and Genetics* 2002; **48**(4):611–617.

[2] Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 2004; **56**(2):211–21.

[3] Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Genetics* 1994; **18**(4):309–317.

[4] Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 1997; **271**(4):511–523.

[5] Korber BT, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci U S A* 1993; **90**(15):7176–80.

[6] Clarke ND. Covariation of residues in the homeodomain sequence family. *Protein Sci* 1995; **4**(11):2269–78.

[7] Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bhlh protein domains: an information theoretic analysis. *Mol Biol Evol* 2000; **17**(1):164–78.

[8] Tillier ER, Lui TW. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 2003; **19**(6):750–755. 1367-4803.

[9] Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 2005; **21**(22):4116–24.

[10] Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 2008; **24**(3):333–340.

[11] Lee BC, Kim D. A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics* 2009; **25**(19):2506–2513.

[12] Merkl R, Zwick M. H2r: identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. *BMC Bioinformatics* 2008; **9**:151.

[13] Lockless S, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 1999; **286**(5438):295–299.

[14] Suel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 2003; **10**(1):59–69.

[15] Najeeb Halabi SL Olivier Rivoire, Ranganathan R. Protein sectors: Evolutionary units of three-dimensional structure. *Cell* 2009; **138**:774–786.

[16] Dekker J, Fodor A, Aldrich R, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics* 2004; **20**(10):1565–1572.

[17] Bickel P, Kechris K, Spector P, Wedemayer G, Glazer A. Finding important sites in protein sequences. *Proceedings of the National Academy of Sciences of the United States of America* 2002; **99**(23):14 764–14 771.

[18] Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol* 1999; **287**(1):187–98.

[19] Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R. Detecting coevolving amino acid sites using bayesian mutational mapping. *Bioinformatics* 2005; **21 Suppl 1**:i126–35.

[20] Caporaso G, Smit S, Easton B, Hunter L, Huttley G, Knight R. Detecting coevolution without phylogenetic trees? tree-ignorant metrics of coevolution perform as well as tree-aware metrics. *BMC Evolutionary Biology* 2008; **8**(1):327.

[21] Halperin I, Wolfson H, Nussinov R. Correlated mutations: advances and limitations. a study on fusion proteins and on the cohesin-dockerin families. *Proteins* 2006; **63**(4):832–845.

[22] Horner DS, Pirovano W, Pesole G. Correlated substitution analysis and the prediction of amino acid structural contacts. *Brief Bioinform* January 2008; **9**(1):46–56.

[23] Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT. Rewiring the specificity of two-component signal transduction systems. *Cell* 2008; **133**(6):1043–1054. 1097-4172.

[24] Eyal E, Frenkel-Morgenstern M, Sobolev V, Pietrokovski S. A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. *Proteins* 2007; **67**(1):142–153.

[25] Liu B, Wang X, Lin L, Tang B, Dong Q, Wang X. Prediction of protein binding sites in protein structures using hidden markov support vector machine. *BMC Bioinformatics* 2009; **10**(1):381.

[26] Madaoui H, Guerois R. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proceedings of the National Academy of Sciences* 2008; **105**(22):7708–7713.

[27] Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP. JOY: protein sequence-structure representation and analysis. *Bioinformatics* 1998; **14**(7):617–23.

[28] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995; **247**(4):536–40.

[29] Wang G, Dunbrack J R L. PISCES: a protein sequence culling server. *Bioinformatics* 2003; **19**(12):1589–91.

[30] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000; **28**(1):235–42.

[31] Holm L, Sander C. Parser for protein folding units. *Proteins* 1994; **19**(3):256–268. 0887-3585.

[32] Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein science : a publication of the Protein Society* 1995; **4**(5):872–884. 0961-8368.

[33] Sowdhamini R, Blundell TL. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci* 1995; **4**(3):506–520. 0961-8368.

[34] DeLano WL. The PyMOL molecular graphics system 2008. URL `www.pymol.org`.

[35] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**(3):403–10.

[36] 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):3389–402.

[37] Xu F, Du P, Shen H, Hu H, Wu Q, Xie J, Yu L. Correlated mutation analysis on the catalytic domains of serine/threonine protein kinases. *PLoS ONE* 2009; **4**(6):e5913.

[38] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; **22**(13):1658–9.

[39] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**(5):1792–7.

[40] Gouveia-Oliveira R, Sackett PW, Pedersen AG. MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics* 2007; **8**:312.

[41] Marcou G, Rognan D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling* January 2007; **47**(1):195–207.

[42] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH–a hierarchic classification of protein domain structures. *Structure* 1997; **5**(8):1093–108.

[43] Bilwes AM, Alex LA, Crane BR, Simon MI. Structure of CheA, a signal-transducing histidine kinase. *Cell* 1999; **96**(1):131–41.

[44] Griswold IJ, Zhou H, Matison M, Swanson RV, McIntosh LP, Simon MI, Dahlquist FW. The solution structure and interactions of CheW from Thermotoga maritima. *Nat Struct Biol* 2002; **9**(2):121–125. 1072-8368.

[45] Park SY, Borbat P, Gonzalez-Bonet G, Bhatnagar J, Pollard A, Freed J, Bilwes A, Crane B. Reconstruction of the chemotaxis receptor kinase assembly. *Nature Structural & Molecular Biology* 2006; **13**(5):400–407. 1545-9993.

[46] Hamer R, Chen PY, Armitage JP, Reinert G, Deane CM. Deciphering chemotaxis pathways using cross species comparisons. *BMC Systems Biology* ; **4**(3).

[47] Quezada CM, Gradinaru C, Simon MI, Bilwes AM, Crane BR. Helical shifts generate two distinct conformers in the atomic resolution structure of the CheA phosphotransferase domain from thermotoga maritima. *Journal of molecular biology* 2004; **341**(5):1283–1294.

[48] Usher KC, de la Cruz AF, Dahlquist FW, Swanson RV, Simon MI, Remington SJ. Crystal structures of CheY from thermotoga maritima do not support conventional explanations for the structural basis of enhanced thermostability. *Protein science : a publication of the Protein Society* February 1998; **7**(2):403–412.

[49] Buckland M, Gey F. The relationship between recall and precision. *Journal of the American Society for Information Science* 1994; **45**(1):12–19.

[50] Fawcett T. An introduction to ROC analysis. *Pattern Recogn. Lett.* 2006; **27**(8):861–874.

[51] N Stephens MA. Use of the Kolmogorov-Smirnov, Cramer-Von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society. Series B* 1970; **32**(1):115122.

[52] Sourjik V. Receptor clustering and signal processing in chemotaxis. *Trends in Microbiology* December 2004; **12**(12):569–576.

[53] Swanson RV, Schuster SC, Simon MI. Expression of CheA fragments which define domains encoding kinase, phosphotransfer, and CheY binding activities. *Biochemistry* August 1993; **32**(30):7623–7629.

[54] Zhu X, Amsler CD, Volz K, Matsumura P. Tyrosine 106 of CheY plays an important role in chemotaxis signal transduction in escherichia coli. *J. Bacteriol.* July 1996; **178**(14):4208–4215.

[55] Park SYY, Beel BD, Simon MI, Bilwes AM, Crane BR. In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. *Proceedings of the National Academy of Sciences of the United States of America* 2004; **101**(32):11 646–11 651.

[56] Volz K. Structural conservation in the CheY superfamily. *Biochemistry* November 1993; **32**(44):11 741–11 753.

Figure Legends

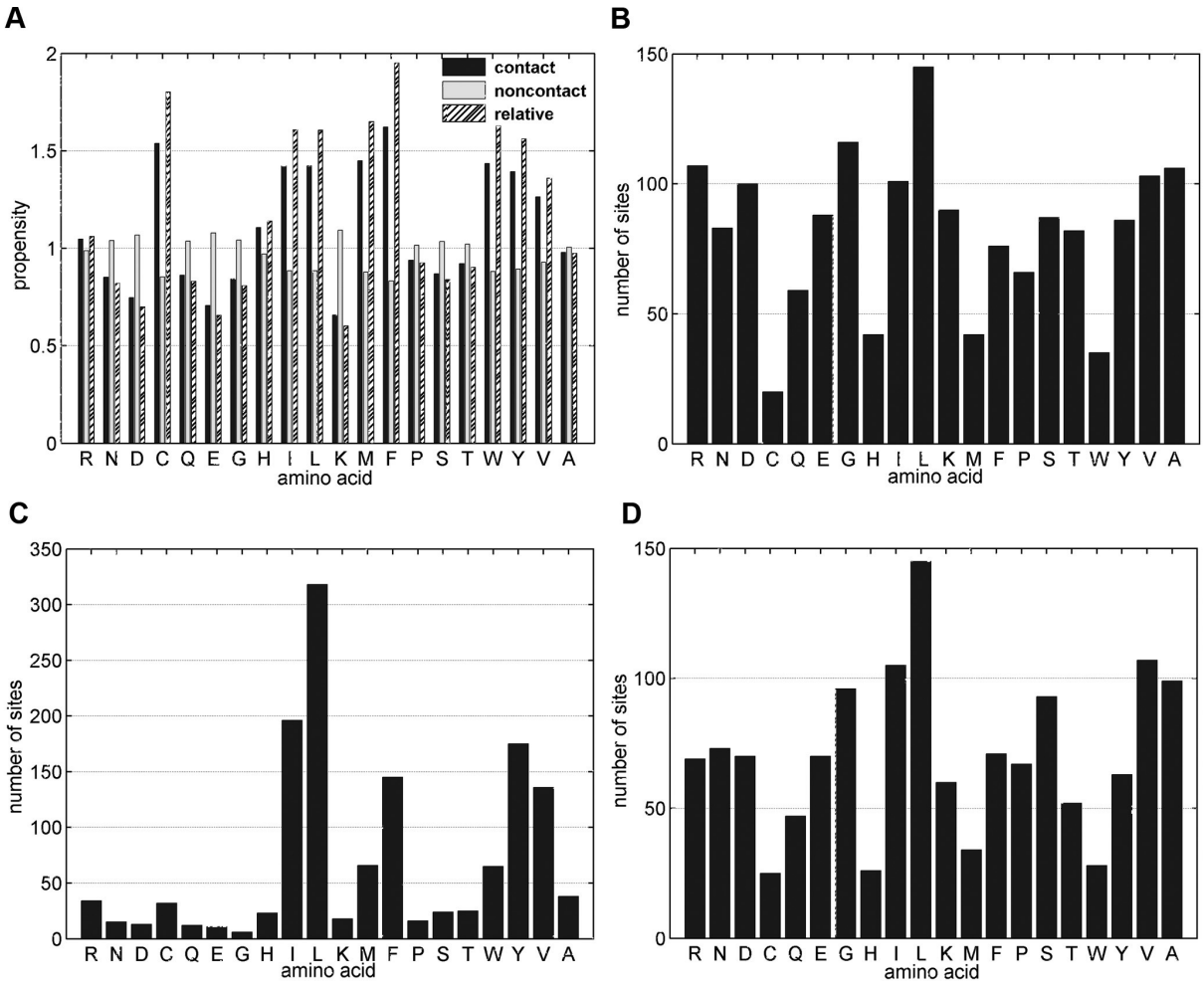


Figure 1: **Occurrence of amino acids in contact sites.** A. Amino acid propensities of each of the 20 amino acids to be in contact sites and non-contact sites, and their relative propensities, calculated using surface-exposed residues in the Propensity data set. Relative propensities are the ratio of contact propensities to non-contact propensities. B. Composition of contact sites in the Fitting data set. C. Composition of amino acids in the top 40 predictions on the Fitting data set given by an amino acid propensity score which does not use patch information (see supplementary material). D. Composition of amino acids in the top 40 predictions on the Fitting data set given by the APro score which does use patch information. The APro predictions have an amino acid composition far more similar to that of real contact sites than the mAPro predictions.

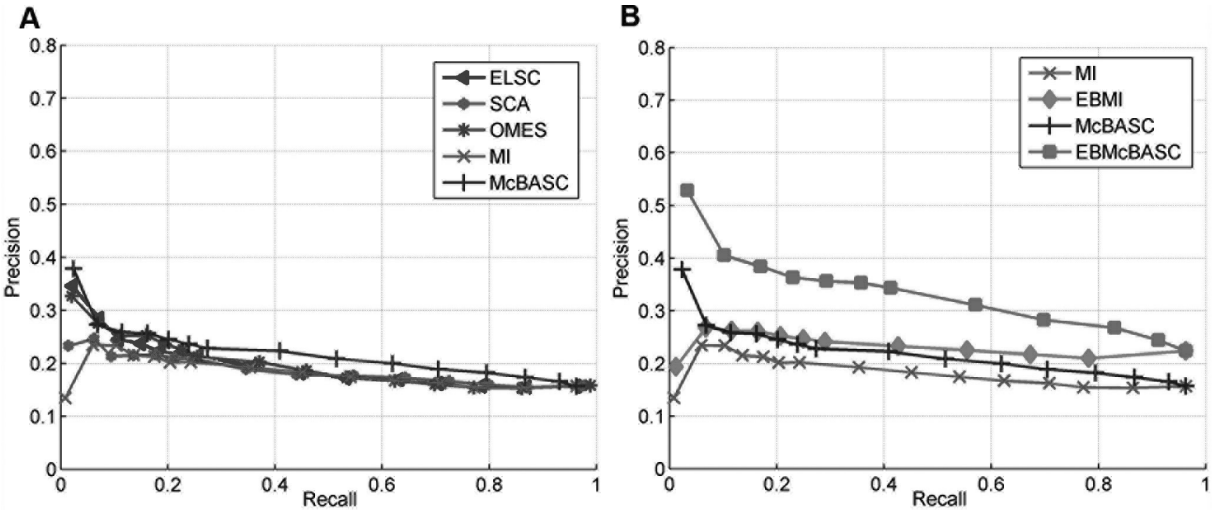


Figure 2: **PROC curves for existing algorithms on the Fitting data set.** A. PROC curves of ELSC, SCA, OMES, MI and McBASC. B. Comparison between the MI and McBASC algorithms with and without using exposed/buried information. Here, any residues which are buried are given a score of 0.

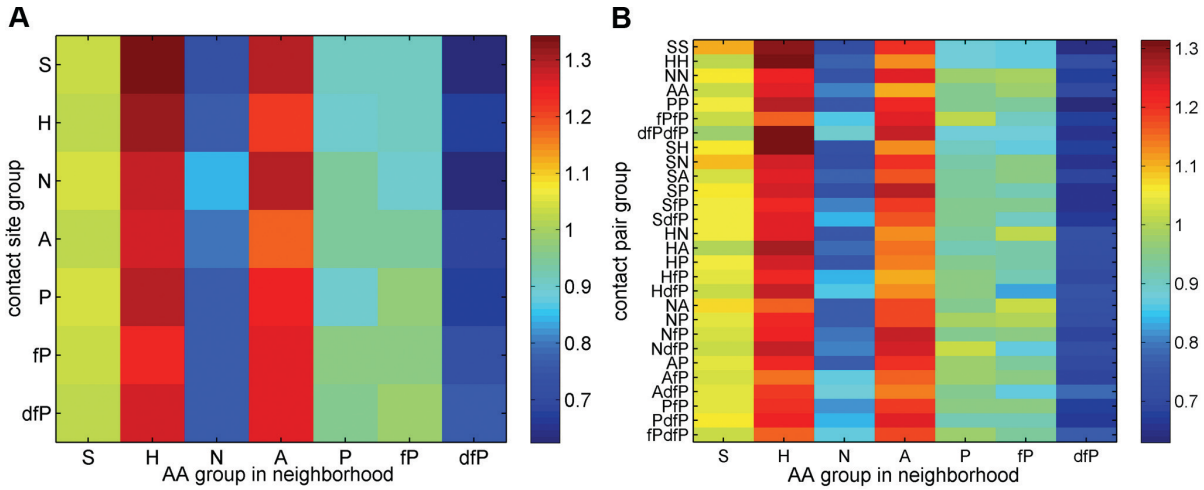
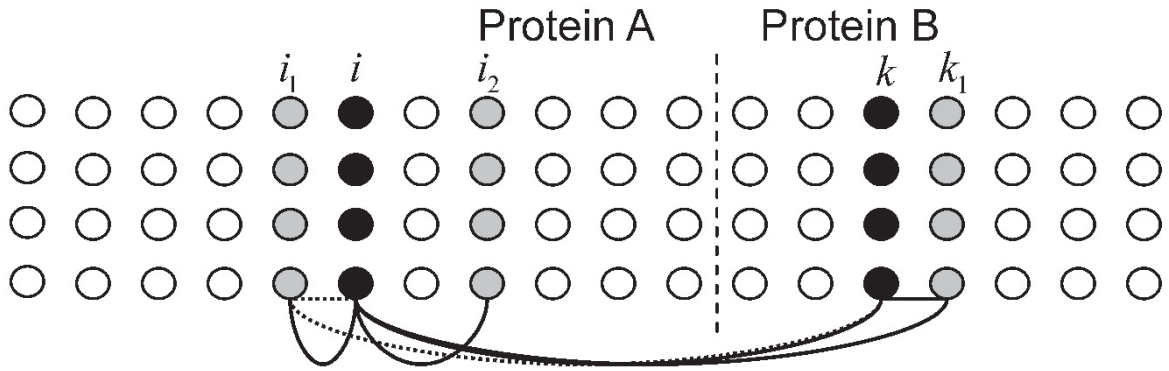


Figure 3: **Heat maps showing propensities of residue types to be neighbours of contact residues.** Residue categories are as follows: S-small, H-hydrophobic, N-negative, A-aromatic, P-polar, fP-favoured postive (R and H) and dfP-disfavoured positive (K). AA=amino acid. A. Intra-domain neighbour propensities for single residue types. For example, if a contact residue is hydrophobic (H), there is a high propensity that a neighbouring residue (within 4.5Å in the same domain) will also be hydrophobic (H), and a low propensity that a neighbouring residue will be negatively charged. B. Intra-domain neighbour propensities for pairs of residue types. For example, if a pair of residues (one from each domain) are in contact, and both residues are disfavoured positively charged (the inter-domain contact pair dfPdfP), there is a high propensity that a neighbouring residue in one or other of the domains is aromatic (A).



$$S_i^{\mathrm{TPro}} = \frac{1}{|\Pi(i)|} \sum_{i_t \in \Pi(i)} w_{ii_t}^{\mathrm{intra}} S_{i_t}^{Triangle},$$

$$\text{where } w_{ii_t}^{\mathrm{intra}} = \frac{1}{M - |G(i_t) \cup G(i)|} \sum_{j \in G(i_t)^c \cup G(i)^c} w^{\mathrm{intra}}(C_{ji_t} \mid C_{ji}).$$

Figure 4: **Calculation of the triangle propensity score (TPro).** The S_i^{TPro} score is calculated for each surface-exposed site i on protein A. We define triangles as three surface-exposed residues from two different proteins, where two of the residues are present on the same protein, less than 4.5Å apart from each other, and a third residue is on the surface of the other protein. Intra-domain weights are calculated using surface-exposed residues in the same domain as site i which are less than 4.5Å away from site i.

Table 1: **Data sets for amino acid propensities**

Data set	Proteins	Contact pairs	Contact residues
Domain	1122 (1150 chains)	83963	65128
Complex	677	49929	40413
Total	1799	133892	105541

Contact residues are those which are used in all the contact pairs. A contact pair may contain a residue which is also present in one or more other contact pairs.

Table 2: **Similarity between amino acid propensities**

Propensity	Amino acid	Pair	Triangle
p-value	0.9568	0.9654	0.9950

Propensities calculated from the domain-domain and protein complex data sets were compared using a two-sample Kolmogorov-Smirnov test [51]. The null hypothesis is that these two samples come from the same distribution. This hypothesis is rejected if the test is significant at the 5% level, *i.e.* the p-value of this test is less than 0.05.

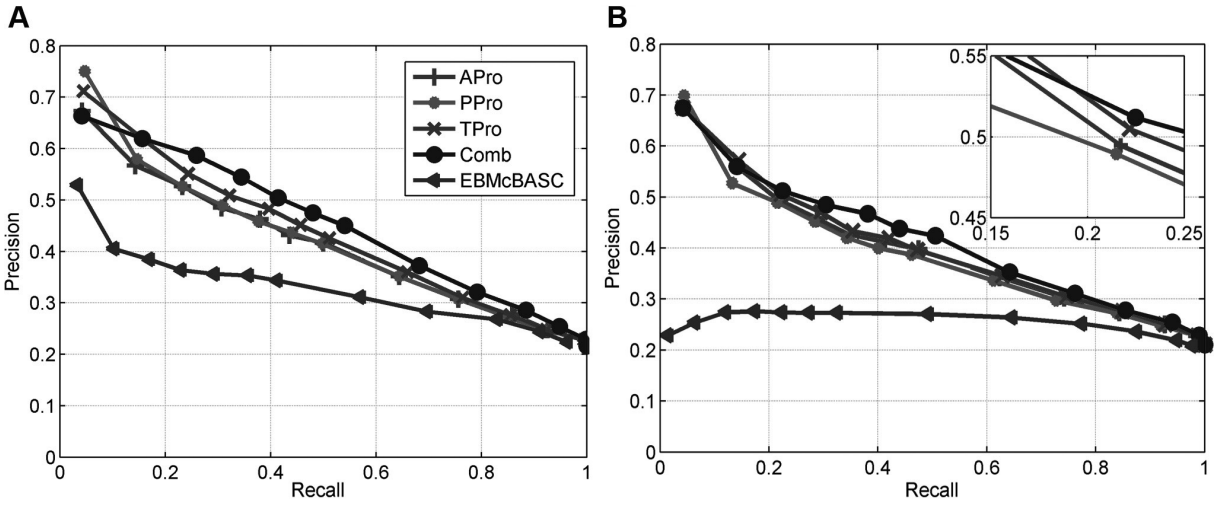


Figure 5: **PROC curves comparing the different i-Patch scores and EBMcBASC.** A. Results from the Fitting data set B. Results from the Blind data set. EBMcBASC was selected as it performed best out of the five correlated mutation algorithms on the Fitting data set. All i-Patch scores significantly outperform EBMcBASC on both data sets.

Table 3: **Data sets for testing of i-Patch scores**

Data set	Proteins	Contact pairs			Contact sites		
		Total	Deleted*	Remaining	Total	Deleted*	Remaining
Fitting	31†	2103	83	2020	1690	38	1652
Blind	31	1740	86	1654	1362	41	1321

*Columns in the final alignment with more than 50% gaps are deleted before i-Patch is run. By doing this, a small number of contact sites will be deleted. †35 domain-domain interfaces are present in the Fitting data set.

Table 4: **Coefficients for the combined model**

Comb	Constant	APro	PPro	TPro	EBMcBASC
Coefficient	-5.5267	3.2675	-3.1180	3.0213	2.1648
p-value	0.0000	0.0000	0.0007	0.0000	0.0000

The coefficients and p-values of the logistic model, Comb, are given by fitting this model on the Fitting data set, using the generalized linear model fitting tool 'glmfit' in Matlab. If a coefficient has a p-value of less than 0.05, it is significantly different from zero.

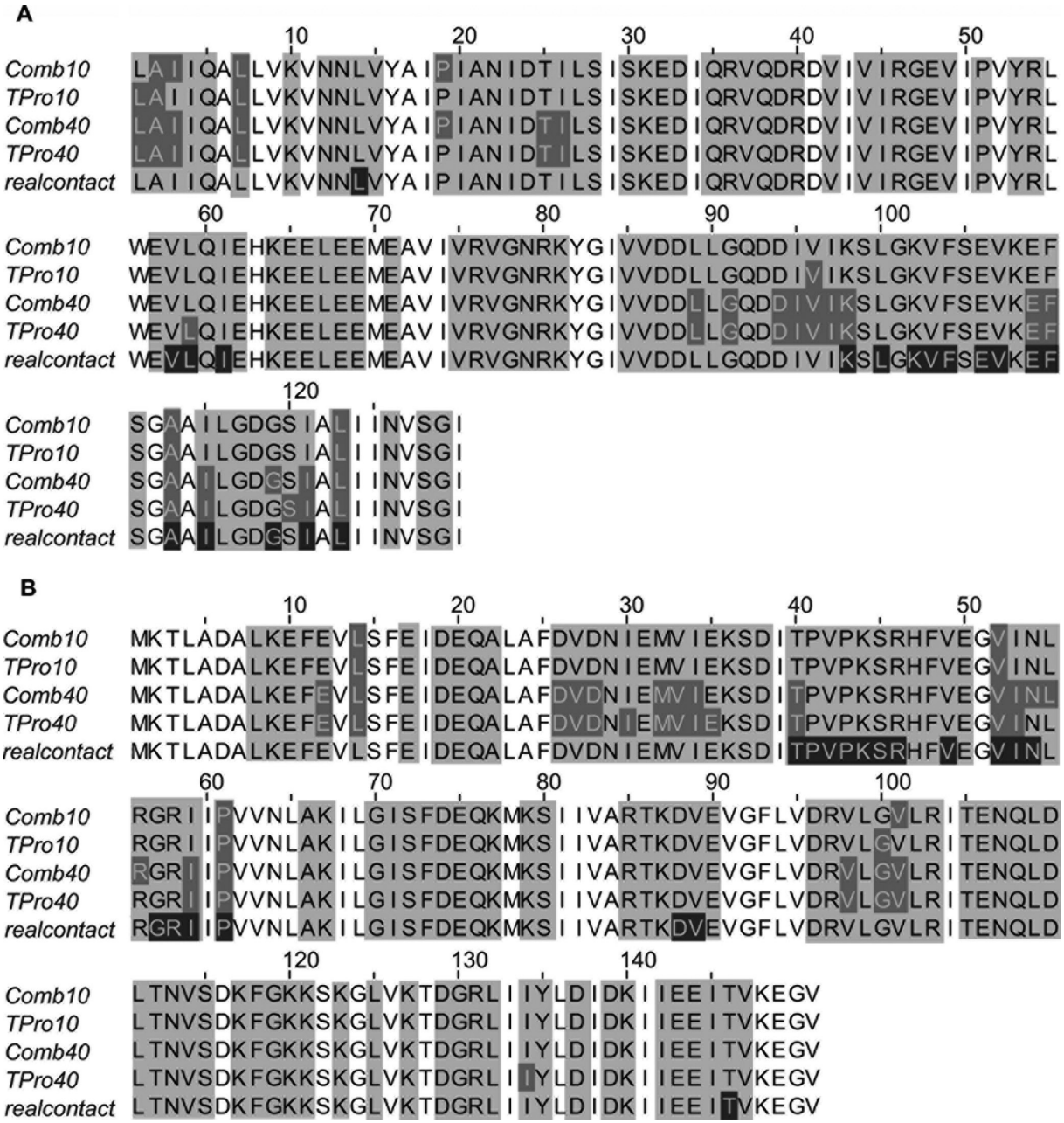


Figure 6: **Predictions given by i-Patch for the complex of CheA and CheW.** A. Predictions on CheAP5. B. Predictions on CheW. The PDB structure 2CH4 is a complex of the chemotaxis proteins CheA (P4 and P5 domains) and CheW from *Thermotoga maritima*. Only the P5 domain of CheA was used in our analysis as this is the domain known to bind to CheW. The top 10, then the top 40, predictions given by TPro and Comb are shown on lines 1-4. The real inter-protein contact sites are highlighted in black on the last line. Surface-exposed regions are shaded.

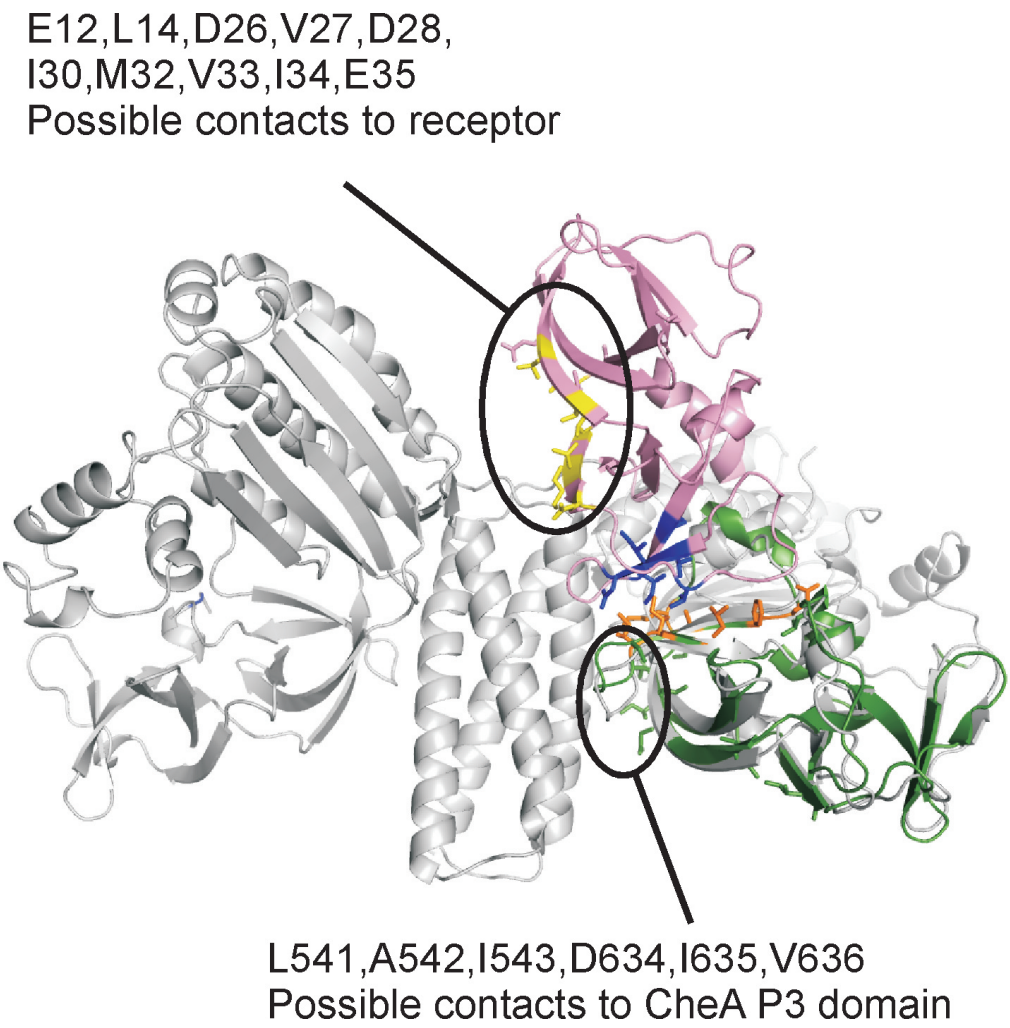


Figure 7: **Predictions given by i-Patch for the CheA-CheW complex, shown on the 2CH4 structure.** 2CH4 is the complex of CheA P4 and P5 domains (only P5 is shown, green) with CheW (pink) from *Thermotoga maritima*. Here this structure is superimposed onto 1B3Q, the structure of CheA P3, P4 and P5 domains from *Thermotoga maritma* (grey). The top 40 sites predicted by TPro are shown as sticks, with those overlapping with real contact sites between CheAP5 and CheW shown in orange and blue. Some sites predicted by TPro, but which are not at the interface between CheAP5 and CheW, may be contacting other CheA domains, or possibly the tip of a chemotaxis receptor. Predicted sites which overlap with mutations that affect the receptor Tar binding in *E. coli* CheW are shown in yellow.

Figure 8: **Predictions given by i-Patch for CheA3 P1 domain and CheY6.** Structures of CheA P1 and CheY from *Thermotoga maritima* (PDB identifiers 1TQG [47] and 1TMY [48] respectively) were used as reference structures for these predictions. A. Predictions on CheA3 P1 domain. B. Predictions on CheY6. Both sequences are those from the *Rhodobacter sphaeroides* proteins. The top 10, then the top 40, predictions given by TPro and Comb are shown on lines 1-4. The surface-exposed regions are shaded.

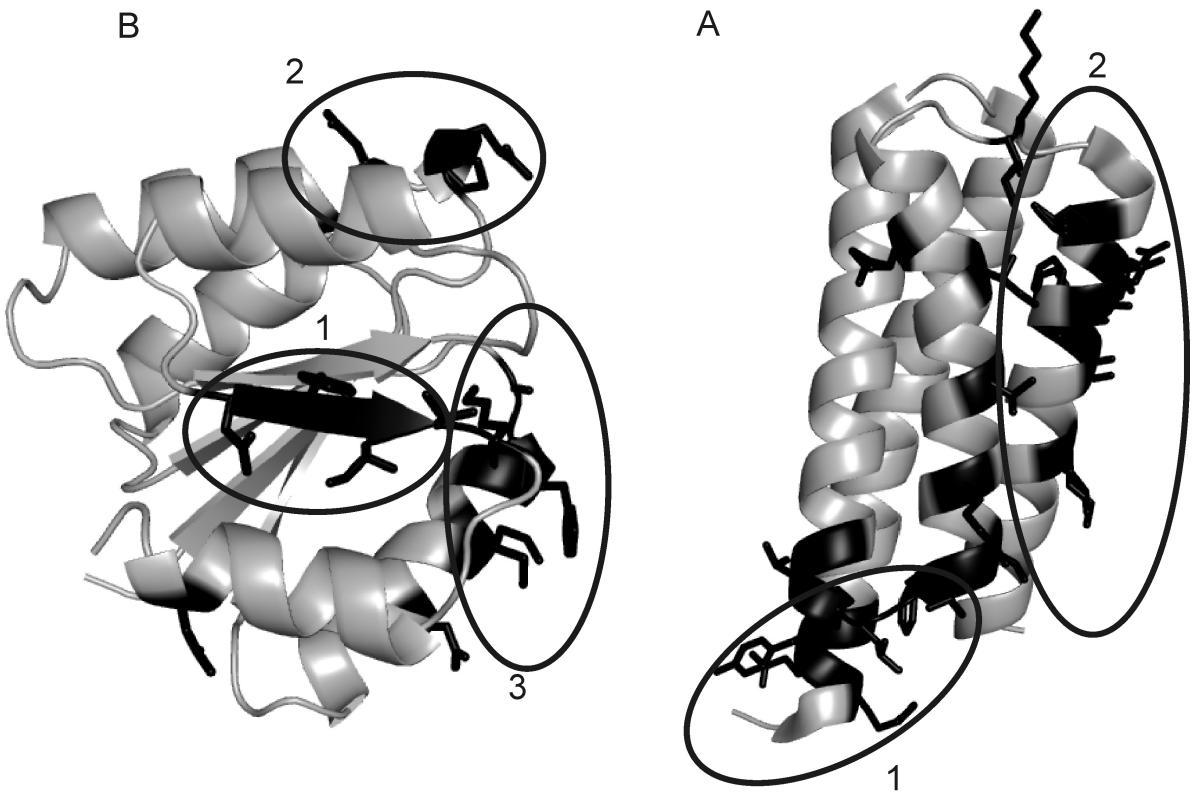


Figure 9: **Predictions given by i-Patch for CheA and CheY shown on reference structures.** A. TPro contact site predictions shown as black sticks on the reference structure CheA P1 from *Thermotoga maritima* (PDB identifier 1TQG [47]). B. TPro contact site predictions shown as black sticks on the structure of CheY from *Thermotoga maritima* (PDB identifier 1TMY [48]). Numbered patches are described in detail in the main text.