Algorithms **2009**, *2*, 764-789; doi:10.3390/a2020764

OPEN ACCESS

algorithms

ISSN 1999-4893

www.mdpi.com/journal/algorithms

Article

SDPhound, a Mutual Information-Based Method to Investigate Specificity-Determining Positions

Sara Bonella [1,†,‡], **Walter Rocchia** [1,2,†,⋆], **Pietro Amat** [1], **Riccardo Nifosí** [1] **and Valentina Tozzini** [1]

[1] NEST Scuola Normale Superiore and CNR-INFM, Piazza dei Cavalieri 7, I-56126 Pisa, Italy

[2] Drug Discovery and Development, Italian Institute of Technology, via Morego 30, Genova I-161631, Italy

[†] These authors contributed equally to the work.

[‡] Current address: Dipartimento di Fisica, Università di Roma ”La Sapienza”, I-00158 Rome, Italy.

E-mails: Sara.Bonella@roma1.infn.it; p.amat@sns.it; r.nifosi@sns.it; v.tozzini@sns.it

[⋆] Author to whom correspondence should be addressed; E-mail: walter.rocchia@iit.it

Received: 22 December 2008; in revised form: 12 March 2009 / Accepted: 5 May 2009 / Published: 26 May 2009

Abstract: Considerable importance in molecular biophysics is attached to influencing by mutagenesis the specific properties of a protein family. The working hypothesis is that mutating residues at few selected positions can affect specificity. Statistical analysis of homologue sequences can identify putative specificity determining positions (SDPs) and help to shed some light on the peculiarities underlying their functional role. In this work, we present an approach to identify such positions inspired by state of the art mutual information-based SDP prediction methods. The algorithm based on this approach provides a systematic procedure to point at the relevant physical characteristics of putative SPDs and can investigate the effects of correlated mutations. The method is tested on two standard benchmarks in the field and further validated in the context of a biologically interesting problem: the multimerization of the Intrinsically Fluorescent Proteins (IFP).

Keywords: specificity determining positions; intrinsically fluorescent proteins; mutual information.

1. Introduction

Considerable efforts in current biophysical research are devoted to identifying viable procedures to engineer mutations in a protein so as to manipulate its properties in desired directions. Most available methods rely on the assumption that the functional, or structural, specificities among homologs depend on relatively few crucial residues that are conserved among proteins sharing the same feature. The problem thus becomes recognizing such residues. The combinatorial number of possible mutations even in relatively small proteins often makes purely experimental approaches, such as random mutagenesis, not affordable and it compels to create reliable and efficient computational tools to assist experiments by predicting which residues are more likely to affect the desired property.

Several *in silico* approaches for predicting sites with some functional importance in proteins are available [1]. Some need the sequence and some need the knowledge of both sequence and structure. Of particular relevance in this context is a family of methods based on information theory, that build upon the Maximum Likelihood Estimator scheme [2]. These techniques exploit the growing amount of protein sequence data available for a wide variety of organisms. The general strategy consists in performing a statistical analysis of a multiple sequence alignment of proteins of a certain family and in relying on the assumption that since mutations at SDPs change the function of the protein, they are generally conserved between proteins with the same function, but tend to be distinct for proteins with different functions [3]. The correlation between the presence of a residue at a given position in the alignment and the inclusion of the protein in a class, identified by a specific function or quaternary structure, is evaluated through the joint probability $P_p(\alpha, i)$ for the event "amino acid α occurs at position p and the protein belongs to the i^{th} class". The figure of merit in the analysis is provided by the average mutual information:

$$I_p = \sum_{i=1}^{N} \sum_{\alpha=1}^{20} P_p(\alpha, i) \ln \frac{P_p(\alpha, i)}{P_p(\alpha) P(i)} \tag{1}$$

The sum over i spans the number of specificity classes and α covers the amino acid set. $P_p(\alpha)$ is the probability to find residue α at position p and $P(i)$ the probability that a sequence belongs to the i^{th} class. The definition above establishes a measure of the specificity content of a position in the alignment: larger values of I_p are expected to indicate more relevant positions in identifying a given subfamily. Unfortunately, actual values of the probabilities in eq. (1) are not known. They can only be estimated and the main difference among mutual information-based methods, lies in the choice of the most appropriate estimator for the joint probability.

In this work, we describe a protocol, called SDPhound, to identify SDPs and analyze the physical characteristics that may be responsible for their role. The core of the procedure is a prescription for constructing an estimator of $P_p(\alpha, i)$ that is rather general and rigorous in its probabilistic interpretation. Once this estimator is available, the specificity content of the different positions can be ranked via the mutual information. Although the formal structure of the estimator is fixed, variations on the specific functions used in it make it possible to implement a set of steps that, appropriately combined, allow to first screen and then characterize putative SDPs. The steps will be described in detail in sections 2. and 3., a preliminary summary is as follows. The estimator of $P_p(\alpha, i)$ is based on the frequency of appearance of a residue at a given position in a given alignment. We combine this basic ingredient with different smoothing functions that can dress the estimated probability to account for non relevant statistical fluctuations. A reasonable stability of positions ranking with respect to the choice of the smoothing functions is a good indicator of the robustness of the predictions with respect to, for example, bias due to the finite size of the aligned set. The structure of the estimator can also be exploited, together with the concept of substitution classes, to investigate the physical and/or chemical factors responsible for specificity both by an *a priori* classification according to some predefined properties and by an automatically generated partitioning that preserves the mutual information content. The procedure can also be generalized to describe possible pairwise correlation effects among SDPs. The performance of SDPhound is assessed in three applications, as illustrated in sections 4. and 5. We begin by validating the approach by comparing its performance to that of a popular alternative scheme due to Kalinina *et al.* [2] in the applications they chose as benchmarks [2]. In particular, we investigate positions responsible for specific features of the Major Intrinsic Protein (MIP) and the bacterial transcription factor LacI families. In both cases, after verifying that SDPhound performs as well as the more established method in ranking putative SDPs, we shall refine the analysis by examining the physical characteristics of the positions. This second step is, to the best of our knowledge, an original feature of SDPhound. We shall further apply the method to identify a set of mutations able to affect the multimeric state of the Intrinsically Fluorescent Protein (IFP) family. This is a problem of considerable interest since IFP are extremely important in molecular biology and biotechnology where they are used in a variety of *in vivo* visualization techniques [4–6]. The first discovered member of the family was the Green Fluorescent Protein (GFP) from the jellyfish Æquorea Victoria (avGFP) [7]. It has been proved that IFP are almost ideal tags for confocal microscopy and that they can be genetically fused to other proteins and expressed in living cells or organisms without disturbing their physiology [5]. For this reason, in the last 15 years keen attention was devoted to the GFP technology, leading to several dozens avGFP mutants with improved photostability and different optical properties [6]. Moreover, it was recently recognized that homologs of this protein exist in a variety of different sea animals belonging to *Cnidaria* and to *Bilatera* [8, 9]. Although the average identity within the family is only 40%, all of them share the same tertiary fold. The IFP quaternary structure is particularly interesting: the wild-type (WT) proteins exist in nature as tetramers, dimers or, rarely, monomers. A clear characterization of the biological role of multimerization for the IFP is lacking. In Molecular Biology applications, however, the monomeric form is desirable in order to produce small fluorescent probes that can easily be transfected into a cell. More than a hundred fluorescent proteins (WT, homologs and mutants) are known and their structural properties are intensively studied and relatively well understood. These proteins therefore are very well suited to be studied with statistical techniques for the design of mutants with specific properties. In this work, we specialize our method to analyze the positions responsible for transforming a multimeric fluorescent protein in a monomeric one. We also look for mutations able to split a tetramer in two dimers. The predicted mutations compare well with those already recognized as effective by random mutagenesis. Furthermore, a new position, deserving experimental verification, is indicated and independently validated via the MMPBSA technique [10]. After the discussion of the results for the IFP family, conclusions and acknowledgements close the paper.

2. Approach

In the following, we summarize the sequence of steps performed in a typical application of SDPhound, while the precise definition of the various theoretical quantities associated with these steps is postponed to the next section. As stated in the introduction, the ultimate goal of the algorithm we propose is to rank putative specificity determining positions based on their mutual information score. To that end, prior to the application of the SDPhound protocol, a set of homologous proteins selected from the literature or public databases is divided into classes whose members share similar specificity. These homologs undergo multiple alignment and the alignment is then used as an external input in the software implementation of SDPhound to determine the mutual information content of each position. Calculating the mutual information requires to define an estimator for the joint probability in eq. (1). We introduce different realizations of a general prescription for the structure of such probability. This prescription, which contains no tunable parameters, can refer to single positions of the alignment, or consider pairwise correlation among them. The probability estimator can weigh the frequencies of occurrence of amino-acids at given positions in any class by using a scheme that employs BLOSUM substitution matrices [11] (see next section for details). This accounts for the presence of "similar" amino acids and it mitigates the effects of finite size of the available statistical sample and of background similarity of homologs on the frequency-based probability estimation. To further refine the assessment of the statistical significance of ranking, a well known problem for any mutual-information method [12], we use a Z-score criterion described at the end of next section.

Once the most interesting positions are selected, the focus is shifted on identifying the relevant physical characteristics associated with them. To this aim, the concept of residue substitution classes [13] is introduced by grouping the various amino acids in sets, "pigeonholes", identified on the basis of some predetermined properties (e. g. hydrophobicity, charge, size). High ranking positions in these "observable"-based runs are expected to be sensitive to the corresponding property, rather than to residue identity. Furthermore, to explore new properties that were not considered by the manual assignment to the pigeonholes, an automated procedure that maximizes the mutual information content of the high ranking positions with respect to pigeonhole content is outlined. As it is shown in the Supplementary Information (SI), the software implementation of our method, SDPhound, automatically provides a pictorial representation of the results, which are reported in the form of Microsoft Excel worksheets and html pages. The MATLAB code for SDPhound is available at http://homepage.sns.it/∼rocchia/

3. Methods

3.1. The estimator for the probability

The key step in setting up the mutual information for a given problem is the definition of the joint probability for the events whose correlation one is trying to establish. In this section, we introduce one such probability so as to make more rigorous the prescription given by Kalinina and co-workers [2] while retaining the single amino acid substitution scheme they advocate. Moreover, we generalize the joint probability in two ways: (1) by considering concerted substitutions of more than one amino acid at a time; (2) by introducing substitution classes ("pigeonholes") that correspond to specific properties of the amino acids and inferring from the presence of a residue belonging to a given pigeonhole at a given position the physical properties conferring specificity.

3.2. Probability estimator for ranking of individual positions

The estimator of the the joint probability in eq. (1) is defined as

$$P_p(\alpha, i) \approx \tilde{P}_p(\alpha, i) \triangleq \sum_{\beta=1}^{24} f_p(\beta, i) m(\beta \to \alpha), \alpha = 1..20 \tag{2}$$

The equation above can be read as follows: the total probability to find amino acid α at position p is given by the probability, approximated by $f_p(\beta, i)$, to find an amino acid β at that position times the probability, $m(\beta \to \alpha)$, that α substitutes β. Such a function thus describes the event "a given amino acid, α, or a similar one, β, occurs at position p and the protein belongs to class i". The sum extends to 24 due to the possible presence in the alignment of non standard symbols such as B, Z, X, and of the gap symbol "$-$". In our applications, the symbols B, Z, and X are included in the definition of the probability above but they are left out of the space of the events since they bring in redundant information (i.e. we are only including mutually exclusive events).

In actual calculations, $f_p(\beta, i)$ is the relative frequency of occurrence of β in position p of a protein belonging to class i. Bias of the overall alignment can be treated via a weighing procedure as suggested in [15]. The substitution probability $m(\beta \to \alpha)$ is introduced with the intent to smooth the bare relative frequency of occurrence of amino acid α at position p by taking into account residue substitutions occurring at that position weighted according to their similarity. This quantity can be defined in different ways. A natural choice, given its interpretation, is

$$m(\beta \to \alpha) = \frac{q_{\beta,\alpha}}{\sum_\lambda q_{\beta,\lambda}} \tag{3}$$

$q_{\beta,\alpha}$ is closely related, in a sense that will be clear shortly, to the Clustered Target Frequencies matrix provided by the Blocks WWW Server, created by S. Henikoff. It represents the probability of occurrence of two amino acids α and β in the same column of a prototypical block alignment [11]. The importance of a smoothing scheme has been recognized and used in the past, most notably in SDPpred, the approach that we choose to benchmark the performance of our method. In this previous work, however, the substitution matrix is used in the log odds form and contains an *ad hoc* parameter [2]. Our suggestion is parameter free and enables the right hand side of eq. 3 to be interpreted as an actual probability. Gaps, "amino acid" number 24 in the sum above, are treated according to the following prescriptions: a column in the alignment is neglected if more than 30% of its constituents are gaps. In the remaining cases, gaps are considered as an additional amino acid with substitution probabilities obtained by suitably enlarging and rescaling $q_{\beta,\alpha}$. Namely, one row and column are added to this matrix, in such a way that the, fictitious, substitution probabilities toward gaps are proportional to the overall percentage of gaps in the alignment. Values of the substitution probability are determined accounting for the individual propensity (or resistance) of each amino acid to mutate, which is represented by the diagonal elements of the $q_{\beta,\alpha}$ matrix. Finally, the extended matrix thus obtained is suitably normalized. We preferred this approach to alternatives in the literature, where, for instance, $m(gap \to gap) = 1$ and $m(gap \to \alpha) = m(\alpha \to$

$gap) = 0\ \forall\alpha$, since in this latter case the "gapped" positions present spuriously high levels of mutual information. Note that the above prescriptions ensure that:

$$\sum_i \sum_{\alpha=1}^{20,24} \tilde{P}_p(\alpha, i) = 1 \tag{4}$$

(i.e. the probability to find any amino acid, or a gap, at a given position is one, as it should). Note also that, if we assume:

$$m(\alpha \to \beta) = \delta_{\alpha\beta} \tag{5}$$

the unsmoothened probability estimation $\tilde{P}_p(\alpha, i) = f_p(\alpha, i)$ is recovered. In section 5. we will show the effects of different choices for $m(\alpha \to \beta)$ on our system by employing the identity matrix and the BLOSUM45, BLOSUM62 matrices. These last two matrices are often used in the literature. In addition to these choices, we also introduced a "local" BLOSUM matrix, called Bloc, whose elements are the $q_{\alpha,\beta}$ relative to the specific alignment that is being examined. In the presence of an alignment constructed from a sufficiently numerous family, this matrix may keep into account the family's peculiarities better than the non specialized BLOSUMs.

3.3. Probability estimator for ranking of correlated positions

Considering correlated mutations of pairs of residues at different positions involves a straightforward generalization of the scheme described above. It is in fact sufficient to replace the event "amino acid α occurs at position p" with "amino acid α occurs at position p and amino acid β occurs at position q". Eq. (2) then becomes

$$P_{pq}(\alpha, \beta, i) \approx \tilde{P}_{pq}(\alpha, \beta, i) \triangleq \sum_{\gamma,\delta} f_{pq}(\gamma, \delta, i) m^{(2)}[(\gamma, \delta) \to (\alpha, \beta)] \tag{6}$$

and the mutual information related to the pair position $\{p, q\}$ can be calculated as

$$I^{(2)}_{pq} = \sum_{i=1}^{N} \sum_{\alpha=1}^{20,24} \sum_{\beta=1}^{20,24} \tilde{P}_{pq}(\alpha, \beta, i) \ln \frac{\tilde{P}_{pq}(\alpha, \beta, i)}{\tilde{P}_{pq}(\alpha, \beta) f(i)} \tag{7}$$

To reduce the impact of marginal mutual information contributions to the pair position score, we maximize the following expression:

$$I^{(2)}_{pq} - I_p - I_q. \tag{8}$$

This definition is very closely related to the one of the "triple mutual information" [16] or "mutual interaction" [17].

We then define the "pair substitution probability" as:

$$m^{(2)}[(\gamma, \delta) \to (\alpha, \beta)] = m(\gamma \to \alpha) m(\delta \to \beta) \tag{9}$$

This choice, beside being the simplest, can be justified as follows: the joint probability of finding residue α at position p and residue β at position q in group i can be written as

$$P_{pq}(\alpha, \beta, i) = P_p(\alpha) P_q((\beta, i)|\alpha_p) \tag{10}$$

where $P_q((\beta,i)|\alpha_p)$ is the conditional probability of finding, in group i, β at q given that α was in p. Both probabilities on the right hand side of the equation above can be written in analogy with the definitions introduced previously:

$$\begin{aligned} P_p(\alpha) \approx \tilde{P}_p(\alpha) &\triangleq \sum_{\delta} f_p(\delta) m(\delta \to \alpha) \\ P_q((\beta,i)|\alpha_p) \approx \tilde{P}_q((\beta,i)|\alpha_p) &\triangleq \sum_{\gamma} f_q((\gamma,i)|\delta_p) m(\gamma \to \beta) \end{aligned} \tag{11}$$

(The use of Henikoff-like matrix element in the definitions above is justified since each one of them refers to single amino acid counts.) Finally, noting that the sample estimate of eq. (10) also leads to

$$f_{pq}(\delta,\gamma,i) \approx f_p(\delta) f_q((\gamma,i)|\delta_p) \tag{12}$$

we found the choice of eq. (9) consistent and reasonable.

3.4. *Probability estimator for ranking based on physical properties of the amino acids*

The concept of substitution class [13] can be used to reduce the alphabet of symbols in our alignment and it allows to capture general features required at a given position to maintain specificity. This point is particularly relevant in our work, for its practical implications. In fact, once a SPD is found, determining the chemical-physical property mainly responsible for its specificity can give hints on the mutations to be tried to push the protein toward another class. This is indeed one of the main application we envision for our method. Additionally, the concept of reducing the 20-letter amino acid alphabet to a few letter one is very general in the theoretical modeling of protein structure, especially in Coarse Grained representations[14]. As illustrated in the following subsections, this reduction can be done either based on a set of physical or chemical characteristics identified *a priori* or by letting SDPhound find the most appropriate set of substitution classes for a given alignment.

A priori partitioning of amino acids on a physical basis

Suppose we characterize and group the amino acids based on some convenient physical property such as size, charge, hydrophobicity, etc. as detailed in the SI. Let σ_α indicate one of these classes, that we shall call "*pigeonholes*" so as to avoid confusion with the specificity classes. Following the same line of thought of subsection 3.2., we define

$$P_p(\sigma_\alpha, i) \approx \sum_{\sigma_\beta} f_p(\sigma_\beta, i) M(\sigma_\beta \to \sigma_\alpha) \tag{13}$$

as the probability to find a member of pigeonhole σ_α in position p in the sequence belonging to class i. For example, if σ_α corresponds to "being a positively charged amino acid", this is the probability to find a basic residue in position p for a class i protein. $f_p(\sigma_\beta, i)$ is the relative occurrence frequency at position p of a generic member of the pigeonhole σ_β and $M(\sigma_\beta \to \sigma_\alpha)$ is the substitution probability of any member of pigeonhole σ_β with one belonging to σ_α. We define this probability as:

$$M(\sigma_\beta \to \sigma_\alpha) = \sum_{\alpha \in \sigma_\alpha} \sum_{\beta \in \sigma_\beta} P(\beta|\sigma_\beta) m(\beta \to \alpha) \tag{14}$$

The probability to pick the element β within the pigeonhole σ_β is estimated according to the usual frequentist approach:

$$P(\beta|\sigma_\beta) \approx \frac{\sum_p f_p(\beta)}{\sum_p f_p(\sigma_\beta)} \tag{15}$$

Once the probability (13) is calculated, it can be used to obtain the mutual information

$$I_p = \sum_{i=1}^{N} \sum_{\sigma_\alpha} P_p(\sigma_\alpha, i) \ln \frac{P_p(\sigma_\alpha, i)}{P_p(\sigma_\alpha) P(i)} \tag{16}$$

Given the meaning of the probability employed, a ranking of the positions based on the mutual information above rewards positions depending on the physical characteristic associated to the pigeonhole. This generalization of standard mutual information based ranking paves the way for an efficient new way of analysing SDPs, because it is able to translate SDP relevance in terms of physical meaning. As we shall show in the applications, this suggests how to target mutations in a more focused way, alternative to purely random mutagenesis.

Optimal automatic pigeonholing

An amino acid partitioning based on some *a priori* chosen property, like the one introduced in the previous subsection, may not capture the characteristic interplay of phenomena responsible for specificity. Therefore, an automated procedure searching for an optimal amino acid distribution among a predefined number of otherwise arbitrary pigeonholes may be preferable. Here we suggest an algorithm that performs this search according to an iterative prescription that minimizes a suitable objective function, defined as the average mutual information of a prescribed number of best ranking positions of the standard run. An iterative cycle is composed as follows: a first partitioning is created by a random assignment of an approximately equal number of amino acids to each pigeonhole. Then, the target function is evaluated. A new partition is generated by a random displacement of an amino acid from the original pigeonhole to a different one. The new value of the target function is computed and compared with the previous one. If this value has increased, the current partitioning is retained, and used as the first step of a new iteration. Otherwise, the new partitioning is rejected and the iteration starts from the original one. The cycle is repeated until the target function becomes stationary. An *a posteriori* analysis of the obtained partitioning can be made to identify the relevant common feature within each pigeonhole.

3.5. Statistical significance calculation

To evaluate the statistical significance of the estimated mutual information for a given position, i, we follow the standard procedure of comparing it to a reference value $< I_i^{sh} >$. The latter is obtained by randomly shuffling residues within columns in the alignment, calculating the mutual information corresponding to each shuffled set and then averaging this quantity over the number of shuffles. By eliminating any correlation between the position and specificity, this procedure provides a measure of the background similarity of the protein set [18]. The computed mutual information is then ranked according to the zeta score

$$Z_i = (I_i - < I_i^{sh} >)/\sigma(I_i^{sh}) \tag{17}$$

where $\sigma(I_i^{sh})$ is the standard deviation of the shuffled mutual information. We have also explored alternative ranking schemes that enforce the additional hypothesis of linear correlation among the values of the mutual information obtained upon shuffling, such as the background correlation removal suggested in[2, 12]. These have, however, proved less informative than the simple method outlined above. Providing an *a priori* criterion to identify the maximum number of meaningful SDPs based solely on the statistical information gathered from the runs is a delicate point of the statistical analysis. In this work, a Z-scores frequency histogram is built to estimate their distribution. Positions with (high) Z-scores that have probability smaller than 0.1 are retained as putative SDPs. This simple indicator provided, in our preliminary study, more stable and consistent results than alternative methods, such as the Bernoulli estimator suggested by [12].

4. Results I: MIP and LacI families

The performance of SDPhound both in identifying putative specificity determining positions and in providing an indication of the relevant physical characteristics associated with them is evaluated on three different protein families. The first two, the MIP and LacI, are established benchmarks [2, 19, 20] for validating mutual information based methods. In this section we begin by comparing the rankings for the positions obtained with our method to those obtained feeding the same alignments to the SDPpred Web Server [21]. This is a well established tool in the field and, as mentioned before, we choose it as a reference method because it too is based on mutual information and introduces an estimator for the joint probability (different from ours) that allows for non-uniform weight of amino acids substitutions. As shown in the next two subsections, the performance of the two methods is very similar, corroborating the reliability of the new scheme presented in this work. After this test, we move to using SDPhound for investigating the characteristics of a suggested SDP. Validation of the results in this case is more delicate and we resort to an *a posteriori* analysis based on visual inspection and physical arguments. In this case too, SDPhound provides interesting information as we shall discuss in the following. The third case we study is that of the IFP family. For this family, the validation of the performance with respect to positions ranking is on quite solid grounds since it is based in large part on comparison with mutations that are experimentally known to affect the property we shall be investigating. Given the non-standard nature of the test set and the biophysical relevance of the problem, results for this case are presented separately in section 5..

4.1. The MIP family

Members of the MIP family assist the transport of both water and small neutral solutes through the cellular membrane. There are about six MIP subfamilies, the two major being the aquaporins (AQP) and the glycerol-uptake facilitators (GLP). Here, we ask SDPhound to identify SDPs responsible for the distinction between these two large subfamilies and then compare our results to those presented in the Kalinina *et al.* paper. The alignment used as input for the different runs is the same as in references [2, 20], where all the necessary details on the nature of the set, containing 60 members of the family, can be found. In table 1 we compare the ranking of putative specificity determining positions obtained using SDPhound.

Table 1. SDPs inferred from 60 members of the MIP family. The Table shows the 24 best ranking SDPs responsible for the distinction between aquaporins and the glycerol-uptake facilitators within MIPs. ($*$) indicate a significant position according to the measure for functional significance used by Kalinina *et al.* [2] (see text for details). The corresponding positions are also colored in red to facilitate the table reading. Results from BLOSUM45 (B45), BLOSUM62 (B62), Identity (Id) and the local BLOSUM (Bloc) substitution matrices are also reported for comparison. Residue numbering refers to GlpF protein from *E. Coli*. In the second row the number of positions reckoned to be significant by the various runs are indicated.

	B45	S	B62	S	Id	S	Bloc	S	SDPpred	S
	21		21		22		24		24	
1	195	*	195	*	207		195	*	207	
2	232		232		187	*	187	*	236	
3	187	*	187	*	232		159	*	48	*
4	186		207		236		108		135	
5	236		108		202	*	232		159	*
6	30		211		48	*	211		187	*
7	108		30		201	*	236		22	
8	207		159	*	159	*	135		195	*
9	211		236		195	*	207		191	*
10	159	*	186		191	*	191	*	201	*
11	48	*	48	*	108		137		108	
12	135		191	*	211		186		137	
13	137		135		137		20		211	
14	194		20		30		30		43	
15	191	*	137		208		48	*	136	
16	20		194		186		134		199	*
17	24		24		20		24		194	
18	26		237		135		216		24	
19	216		199	*	199	*	43		20	
20	237		34		235		202	*	200	*
21	34		132		34		237		193	
22	233		199	*	134		194		194	
23	194		31		26		208		34	
24	100		43		56		34		257	

Columns 1-4 report the outcome of successive runs performed using the different substitution matrices described in section 3.2.: BLOSUM45 (B45), BLOSUM62 (B62), the identity (Id), and the ”local”

BLOSUM matrix (Bloc). The last column reports the results provided by the SDPpred web server. The asterisk in the column marked S in the table indicates a position identified as significant according to the criterion established in [2] to assess the performance of SDPpred, i.e. proximity to the glycerol molecules bound inside the pore channel in the crystal structure 1FX8 (see reference for details). The second row of the table heading indicates the number of statistically significant positions in the ranking based on the Z-score criterion illustrated in section 3.5. (columns 1-4), and by the analogous indicator in the SDPpred scheme (column 5). As it can be seen from the table, both the identity matrix based run of SDPhound and SDPpred identify 8 significant positions. The performance of the SDPhound runs that use the alternative substitution matrices is very similar, and it is in fact difficult to assess if there is a difference given the limited statistics available. Table (2) presents the ranking produced by SDPhound when the mutual information used in the ranking is constructed using the probability defined in eq (13). The method points blindly - i.e. not guided by the researcher's insight into the problem - to a set of positions. Based on the ranking, an *a posteriori* analysis can be performed to evaluate the characteristics of the isolated positions. Among the higher ranking positions, 48 is singled out. The relevance of this site on the functioning of the pore selectivity of the transported molecule has already been remarked experimentally [22, 23]. The residue at this position is part of the bottleneck of the transport channel. In GLPs, the pore cavity is almost exclusively constituted by hydrophobic and few positively charged residues, but in AQPs it is also constituted by some polar ones; in either case, the pore has a slightly different location through the protein. Pigeonholing suggests SDP 48 is important by way of its charge/polarity, while its dimension also has a minor influence.

Table 2. Substitution class based runs over SDPs relevant to aquaporins and the glycerol-uptake facilitators distinction. For each position, the ranking in the run related to hydrophobicity (hyd), charge (crg), and size is shown. "–" means that in that run, the position ranked below 40^{th}. A color code is used to help reading: the positions are colored according to which "quality" (hydrophobicity=red, charge/polarity=green or size=blue) is more relevant to that position. In ambiguous cases "intermediate" colors are used: magenta=blue+red (hyd+size), yellow=red+green (hyd+crg) and cyan=green+blue (crg+size). For practical and representation purposes, we identified a significance threshold of 10 for coloring specific positions. When a position ranks above 10 in all of the three features, it is outlined in bold.

Pos.	195	187	159	48	202	201	191	199	200	207
hyd	2	–	–	–	–	11	16	–	22	–
crg	–	–	–	2	–	3	5	19	–	7
size	3	–	7	14	1	–	13	18	38	2

Pos.	232	236	186	108	135	30	**211**	22	137	20
hyd	5	6	3	12	13	7	4	–	9	–
crg	–	–	–	13	–	–	6	–	14	4
size	6	4	10	8	19	11	5	9	12	15

This is reasonable since, in such a narrow channel, even small changes in polarity and conformation can discriminate between different substrates, altering the balance between the affinity towards a substrate and the efficiency of its transport mechanism. For instance, in some AQPs a more bulky hydrophobic amino acid in this position may completely shut the channel. Two other interesting positions highlighted by the method are 201 and 202. These are located near the bottleneck generated by residue 48. These residues are relevant, respectively, for charge and size. While charge of 201 may modulate substrate affinity, dimension of residue 202 helps explain why the transport channel has different size and position in the GLP and AQP subfamilies.

4.2. The LacI family

As a second test of the usefulness of SDPhound, we apply it to investigating the LacI family, another classic testing ground for these methods. The LacI family is a rather large family of bacterial transcription factors, which comprises 54 orthologous and paralogous proteins divided in 15 subfamilies (AraR, KdgR, CcpA, DegA, YjmH, RbsR, PurR, CytR, GalSR, AscG, LacI, TreT, GntR, IdnR and FruR), populated with a number of members which varies from 2 (for the AraR, KdgR, YjmH, LacI and IdnR subfamilies) to 12 (for the CcpA subfamily). Further details on the nature of the set can be found in the references mentioned above, while here we summarize only the minimal data required for benchmarking SDPhound. The sequence alignment was the same as in [2, 19, 20]. Also for this family, key positions for discriminating among the mentioned subfamilies are searched for. And also in this case, the evaluation of positions indicated by Gelfand and coworkers in [2] is used to compare the performances of our method and of SDPpred. In that work, SDPs were mapped on 3D X-ray structures 1WET, 1JWL and 1BYK and classified into 4 groups, that we indicate with the numbering 1 to 4 in the score columns, marked as S, of Table 3. The classification is as follows: 1) experimentally validated SDPs, 2) SDPs of a domain, with a distance < 5Å from the effector, the substrate or another subunit within a multimer in at least one of the 3D structures and < 7Å in the other two, 3) SDPs in the vicinity of one of the domains, and 4) SDPs with no apparent function. As it can be seen from the Table, this test case too shows equivalence of the new and of the benchmark method. The performances of SDPhound with the identity substitution matrix and of SDPpred are practically equivalent, with close numbers of positions in the four classes indicated above (5 vs. 4 in class 1, 12 vs. 13 in 2, 17 in 3 and only 1 in 4 in SPDhound vs. SDPpred). Positions 73, 190 and 292 are part of the binding site and interact very closely with the substrate. Lu and coworkers [24] have investigated position 190, demonstrating its role in corepressor binding. This position was not predicted by SDPpred but was identified both by the B45 and the Bloc SDPhound runs. When analyzed with the pigeonholing runs, several interesting features of the relevant positions emerge, as summarized in Table 4. The experimentally validated positions (15, 16, 55, and 147) rank quite high in all physical characterization, pointing to the high specificity of the residues at those locations. Position 55 hosts a residue in close contact with operator DNA. The positive charge of K55 in 1WET, as well as other polar residues able to donate a hydrogen bond found in the same position in other ortho/paralogous proteins, explains why it has been found by mutagenesis as a position influencing affinity for the operator DNA [25]. Glasfeld and coworkers demostrated that the positively charged Lys55 is able to discriminate among different operator DNA sequences. Firstly, it increases affinity for the electrostatically negative minor groove of the operator DNA; secondly, it selectively binds operator DNA's that does not contain guanine in the position Lys55 interacts with [26]. Nonpolar residues, such as alanine, appear in other paralogous proteins that can bind guanine-containing operator DNA sequences.

SDPhound not only predicts this position, but it correctly points out that the physical property determining its specificity is based on its charge/polarity. We would like to emphasize that such result is obtained by SPDhound without any structural information of the protein or the interacting DNA. Among the other positions in Table 4, 98 stands out both for charge and size relevance. Inspection of the X-ray structures reveals that residue 98 (or its homologous 100 in PDB structure 1JWL) is located both at the opening of the substrate pocket and in contact with another protein in the dimer. Modulation of size and charge, then, can enforce in different proteins a sophisticated selectivity towards different substrates. Also interesting is the classification of position 107, which is pointed at as relevant due mostly to hydrophobicity: the location of the corresponding residue at the interface could be due to the need of tightening or loosening the bond between the proteins in the complex and, thus, their ability to cooperate.

5. Results II: IFP family

In this section, we show results of the application of the method described above to identify putative SDPs for the multimerization propensity of a set of intrinsically fluorescent proteins. More space is reserved to the description of the runs and of the results performed in this case, since, as described in the introduction, they represent an original set of calculations that are interesting not only as a test case for the method, but also for the biological problem that is addressed. They can also be looked at as a reference protocol for the application of SDPhound to thoroughly investigate SDP's in a family of proteins. In preparation for the calculations, 121 IFP sequences were obtained from the Entrez NCBI public database and classified according to their aggregation state. Out of these, a non-redundant subset of crystallized proteins was extracted and used in a combined structure-based sequence-based alignment procedure as described in 5.1. Section 5.2. presents the results of the standard and pigeonhole analysis for the study of the monomerization in the family. Mutagenesis experiments [27] succeeded in identifying 33 mutations among those that induce monomerization while preserving the fluorescence of the protein. These positions provide a new and quite powerful benchmark, based on experiments, rather than on spatial localization of the residues as for the majority of positions for the MIP and LacI families. We also use this application to show some results obtained with the optimal pigeonholing procedure and the use of the probability correlating the positions of two amino acids. As a final application of SDPhound, in section 5.3. we also consider the problem of identifying positions responsible for the dimeric or tetrameric form of intrinsically fluorescent proteins. In this case, less experimental information is available so we rely on an auxiliary computational tool, the MMPBSA method [10] to validate a particularly interesting mutation suggested by our method. In the SI an Excel file containing the detailed information of which amino acids occurred in each class, for each best ranking position, is provided.

Table 3. SDPs inferred from 54 members of the LacI family. The Table shows the 35 best ranking SDPs that characterize the LacI family of bacterial transcription factors. Column labeling and second row content as in Table 1. Details on the scores S can be found in the text. Positions with different scores are colored in different colors to facilitate reading. Numbering follows *E. Coli* PurR in 1WET PDB structure.

	B45	S	B62	S	Id	S	Bloc	S	SDPpred	S
	32		31		35		33		24	
1	15	1	15	1	160	2	15	1	146	3
2	122	2	122	2	15	1	122	2	160	2
3	144	3	55	1	98	2	160	2	55	1
4	107	3	56	2	55	1	55	1	85	3
5	16	1	144	3	146	3	144	3	98	2
6	55	1	107	3	122	2	249	2	302	3
7	126	3	85	3	114	2	56	2	16	1
8	121	3	98	2	249	2	69	3	221	2
9	56	2	323	3	85	3	85	3	15	1
10	85	3	16	1	147	1	16	1	122	2
11	323	3	126	3	16	1	146	3	114	2
12	123	3	123	3	56	2	302	3	121	3
13	69	3	160	2	221	2	121	3	69	3
14	98	2	249	2	323	3	98	2	50	3
15	146	3	146	3	25	3	147	1	249	2
16	249	2	69	3	91	3	107	3	147	1
17	160	2	121	3	302	3	323	3	56	2
18	20	1	147	1	193	2	233	4	323	3
19	25	3	25	3	157	4	66	3	186	3
20	302	3	156	4	148	3	91	3	144	3
21	156	4	302	3	50	3	50	3	91	3
22	147	1	91	3	133	3	145	3	123	3
23	227	2	20	1	69	3	126	3	53	3
24	21	3	60	2	121	3	25	3	66	3
25	96	2	193	2	144	3	96	2	157	4
26	66	3	227	2	107	3	221	2	57	2
27	50	3	114	2	123	3	156	4	21	3
28	60	2	50	3	192	2	60	2	145	3
29	221	2	221	2	20	1	114	2	61	2
30	91	3	157	4	159	3	294	3	107	3
31	53	3	96	2	110	2	190	1	193	2
32	190	1	294	3	53	3	292	2	29	2
33	73	2	233	4	186	3	123	3	192	2
34	233	4	73	2	227	2	110	2	78	2
35	157	4	66	3	57	2	280	3	133	3

Table 4. Substitution class based runs over SDPs relevant within the LacI family. Ranks and colors as in Table 2.

Pos.	**15**	160	**122**	144	55	98	50	107	56	16
hyd	6	15	5	9	14	–	7	3	17	4
crg	2	8	9	–	4	1	13	17	29	14
size	5	9	7	25	16	1	12	–	4	8
Pos.	**146**	**249**	**126**	**85**	**114**	**121**	**123**	**221**	**147**	**323**
hyd	18	–	1	10	–	–	–	–	–	30
crg	11	18	6	–	15	–	–	–	7	–
size	14	–	–	3	–	2	28	30	17	21

Figure 1. DsRed tetramer (1GGX). Residues relevant for the multimeric properties are shown explicitly, and colored according to the score as in Table 5. The tetramerization is here schematized as *"dimerization of dimers"* (dimer in cyan).

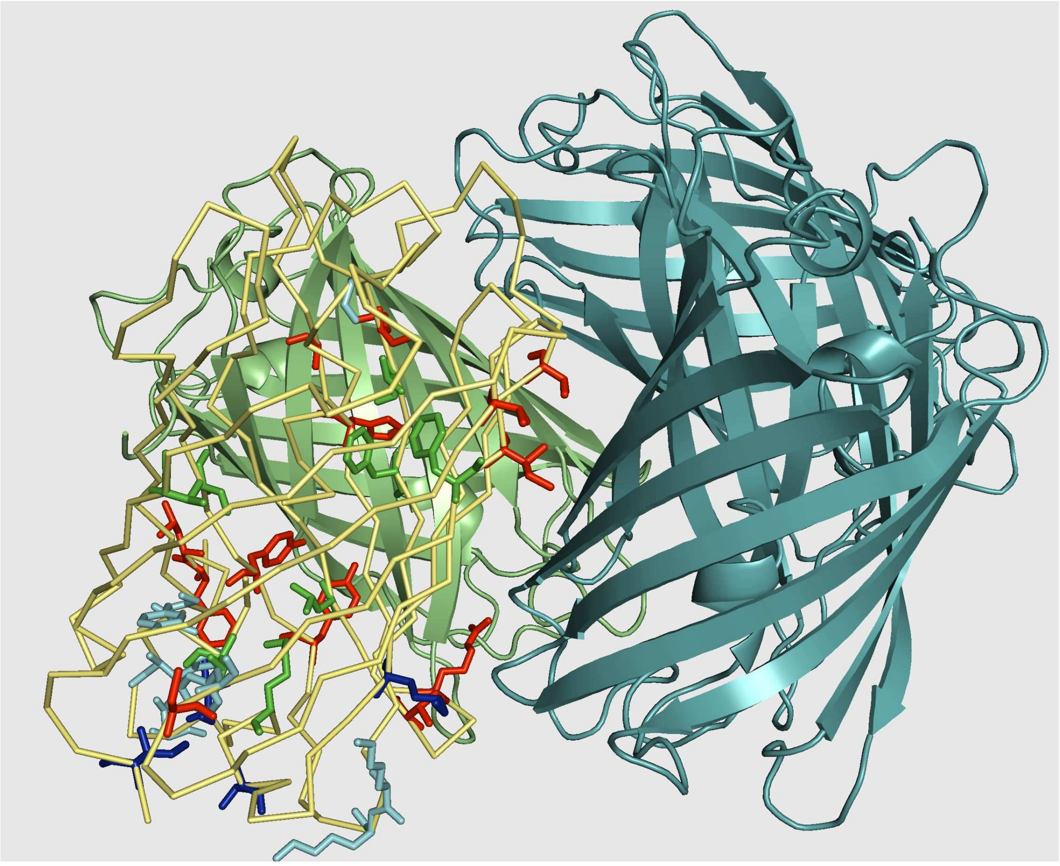


Table 5. The 24 best ranking SDPs inferred from 92 mono- and multimeric IFP members; next 16 SDPs can be found in SI. Column labeling as in Table 1. Numbering refers to the DsRed sequence (1GGX PDB code). The first row reports the number of statistically significant positions. The meaning of the interspersed score columns, S, is detailed in the main text, as well as the description of the peculiar role of position 147. Different scores colored as in Table 1. The last row of the panel reports how many of the 33 experimentally sanctioned positions ranked between 1 and 33 in the corresponding run. Each run, comprising $N = 10000$ shuffles, required about 20 minutes on a 2.13GHz INTEL® PENTIUM® processor with 2GB of RAM.

	B45	S	B62	S	Id	S	Bloc	S	SDPpred	S
	19		24		21		18		9	
1	194	1	194	1	117	1	117	1	117	1
2	117	1	117	1	83	2	194	1	83	2
3	224	1	177	2	194	1	224	1	79	3
4	177	2	224	1	164	1	175	2	194	1
5	164	1	164	1	192	1	177	2	184	4
6	156	1	156	1	156	1	174	1	164	1
7	44	2	192	1	223	1	124	2	185	4
8	197	2	197	2	175	2	192	1	192	1
9	192	1	174	1	153	1	164	1	156	1
10	125	1	44	2	5	1	153	1	175	2
11	174	1	175	2	177	2	162	1	177	2
12	162	1	4	3	224	1	197	2	72	4
13	4	3	125	1	124	2	4	3	153	1
14	175	2	153	1	162	1	83	2	124	2
15	124	2	124	2	147	*	44	2	8	3
16	127	1	162	1	4	3	156	1	147	*
17	153	1	150	2	174	1	71	2	150	2
18	72	4	127	1	72	4	150	2	174	1
19	150	2	92	3	8	3	72	4	44	2
20	92	3	72	3	44	2	78	4	21	1
21	78	4	83	2	21	1	223	1	219	4
22	85	3	85	3	71	2	147	*	162	1
23	6	1	78	4	6	1	125	1	75	4
24	83	2	118	4	197	2	1	4	57	4
	20		22		23		22		18	

5.1. Sequence alignment

The only external input necessary to perform SDPhound runs is a multiple sequence alignment. A common feature of sequence-based bioinformatic methods is the potentially high sensitivity to the quality of the given alignment. Therefore, particular care has to be devoted to maximize its accuracy before attempting the statistical analysis. This can be sometimes a difficult task (as, for example, in the case of receptors with poor structural characterization and varying sequences); the description of what we did in the IFP case to gain confidence in the reliability of the input follows. We adopted a combined structure-based and sequence-based alignment procedure. First we structurally aligned the protein sequences for which 3D structure is available using the "STAMP" multiple alignment module of VMD1.8.3 [28, 29]. This procedure is based on the alignment of the structurally conserved regions, thus it is likely to bring more information and accuracy in the alignment. For each of the proteins with unknown crystal structure, we determined the closest homolog in a non redundant sub-set of the crystallized proteins set. Sequence-based alignment is done with the FASTA suite of programs.

Subsequently, we aligned with a standard multiple-alignment sequence based algorithm all of the proteins with the same crystal homolog. Finally we merged all the sequence-based alignments following the first structure-based alignment to produce the input data. When no crystallographic data are available, this procedure reduces to a standard multiple sequence based alignment. Conversely, when all the sequences correspond to a crystallographic structure, this procedure reduces to a standard structure-based alignment. Figure 2 reports a selection of the alignment. The proteins are shown and the SDPs are outlined with color coding as in Table 5. The whole alignment is given in the SI.

5.2. Monomerization of multimeric IFP

Within the IFP, different monomer to tetramer transitions can occur. Here we focus on the one that is expected to characterize close homologs of the DsRed protein, Figure 1. A set of single position runs aiming at identifying putative SDPs that favor the monomeric character was performed as a first test of the procedure. The sequences were divided in two classes consisting of 26 monomeric and 66 multimeric proteins. Average identities were 51.6%, 73.6% and 49.9% for the whole set, the first and the second class, respectively. Table 5 shows the 20 best ranking positions, estimated according to the different choices of similarity matrix described in the Methods. As in the previous section, for classification and assessment purposes, we associated a numeric code, S, to each position. Consistent with the fact that experimentally validated positions can be categorized as belonging to the multimer forming interface or not, S values are defined as: S=1 experimentally validated, at the multimer-forming interface; S=2 experimentally validated, not at the multimer-forming interface; S=3 untested, at the multimer-forming interface; S=4 untested, not at the multimer-forming interface.

Figure 2. Alignment of a subset of 18 crystallized IFPs, labeled according to their pdb code. The sequences are separated according to their multimeric class and amino acids in specific positions colored as in Table 6 (complete alignment given in the SI). The whole IFP set has an overall average identity of 40.8%. The sequence identity is larger in the first part of the sequence up to position $\sim$100, and smaller in the second half, except for isolated residues such as the highly conserved E197 (numbering according to DsRed).

```
Monomeric
2H50    MVSKGEENNMAIIKEFMRFKVRMEGSVNGHEFEIEGEGEGRPYEGFQTAKLKVTK--GGPLPFAWDILSPQFTYGSKAYVKHPADIP   85
Dimeric
1YZW    --------MAGLLKESMRIKMYMEGTVNGHYFKCEGEGDGNPFTGTQSMRIHVTE--GAPLPFAFDILAPCCEYGSRTFVHHTAEIP   77
2G6X    -------------LPAMEIECRITGTLNGVEFELVGGGEGTPEQGRMTNKMKSTK---GALTFSPYLLSHVMGYGFYHFGTYPSGYE   71
2G6Y    -------------LPAMEIECRITGTLNGVEFELVGGGEGTPEQGRMTNKMKSTK---GALTFSPYLLSHVMGYGFYHFGTYPSGYE   71
Tetrameric
2A48    -----MALSNKFIGDDMKMTYHMDGCVNGHYFTVKGEGNGKPYEGTQTSTFKVTMANGGPLAFSFDILSTVFKYGNRCFTAYPTSMP   82
2A47    -----MALSNKFIGDDMKMTYHMDGCVNGHYFTVKGEGNGKPYEGTQTSTFKVTMANGGPLAFSFDILSTVFKYGNRCFTAYPTSMP   82
1ZGP    -----MRSSKNVIKEFMRFKVRMEGTVNGHEFEIEGEGEGRPYEGHNTVKLKVTK--GGPLPFAWDILSPQFQYGSMVYVKHPADIP   80
1XAE    -----MAHSKHGLKEEMTMKYHMEGCVNGHKFVITGEGIGYPFKGKQTINLCVIE--GGPLPFSEDILSAGFKYGDRIFTEYPQDIV   80
1XA9    -----MAHSKHGLKEEMTMKYHMEGCVNGHKFVITGEGIGYPFKGKQTINLCVIE--GGPLPFSEDILSAGFMYGDRIFTEYPQDIV   80
1XSS    ---------VSVITSEMKMELRMEGAVNGHKFVITGKGSGQPFEGIQNMDLTVIE--GGPLPFAFDILTTVFDYGNRVFVKYPEEIV   76
1ZUX    ---------MSAIKPDMKINLRMEGNVNGHHFVIDGDGTGKPFEGKQSMDLEVKE--GGPLPFAFDILTTAFHYGNRVFAEYPDHIQ   76
1BTJ    ---------MSAIKPDMKINLRMEGNVNGHHFVIDGDGTGKPFEGKQSMDLEVKE--GGPLPFAFDILTTAFHYGNRVFAEYPDHIQ   76
1GGX    -----MRSSKNVIKEFMRFKVRMEGTVNGHEFEIEGEGEGRPYEGHNTVKLKVTK--GGPLPFAWDILSPQFQYGSKVYVKHPADIP   80
1XQM    --------MASFLKKTMPFKTTIEGTVNGHYFKCTGKGEGNPFEGTQEMKIEVIE--GGPLPFAFHILSTSCMYGSKTFIKYVSGIP   77
2A54    --------MASFLKKTMPFKTTIEGTVNGHYFKCTGKGEGNPFEGTQEMKIEVIE--GGPLPFAFHILSTSCMYGSKTFIKYVSGIP   77
1XMZ    --------MASFLKKTMPFKTTIEGTVNGHYFKCTGKGEGNPFEGTQEMKIEVIE--GGPLPFAFHILSTSCMYGSKTFIKYVSGIP   77
1UIS    --------MNSLIKENMRMMVVMEGSVNGYQFKCTGEGDGNPYMGTQTMRIKVVE--GGPLPFAFDILATSFMYGSKTFIKHTKGIP   77
1MOU    ---------MSVIATQMTYKVYMSGTVNGHYFEVEGDGKGRPYEGEQTVKLTVTK--GGPLPFAWDILSPQCQYGSIPFTKYPEDIP   76
            0    .   10    .   20    .   30    .   40    .   50      .   60    .   70    .   80

Monomer
2H50    DYFKLSFP-EGFKWERVMNFEDGGVVTVTQDSSLQD---GEFIYKVKLRGTNFPSDGPVMQKKTMGWEASSERMYPEDG--ALKGEI 166
Dimeric
1YZW    DFFKQSFP-EGFTWERTTTYEDGGILTAHQDTSLEG---NCLIYKVKVLGTNFPADGPVMKNKSGGWEPCTEVVYPENG--VLCGRN 158
2G6X    NPFLHAINNGGYTNTRIEKYEDGGVLHVSFSYRYEA---GRVIGDFKVMGTGFPEDSVIFTDKIIRSNATVEHLHPMGDN-DLDGSF 154
2G6Y    NPFLHAINNGGYTNTRIEKYEDGGVLHVSFSYRYEA---GRVIGDFKVMGTGFPEDSVIFTDKIIRSNATVEHLHPMGDN-DLDGSF 154
Tetrameric
2A48    DYFKQAFP-DGMSYERTFTYEDGGVATASWEISLKG---NCFEHKSTFHGVNFPADGPVMAKKTTGWDPSFQKMTVCDG--ILKGDV 163
2A47    DYFKQAFP-DGMSYERTFTYEDGGVATASWEISLKG---NCFEHKSTFHGVNFPADGPVMAKKTTGWDPSFEKMTVCDG--ILKGDV 163
1ZGP    DYKKLSFP-EGFKWERVMNFEDGGVVTVTQDSSLQD---GCFIYKVKFIGVNFPSDGPVMQKKTMGWEASTERLYPRDG--VLKGEI 161
1XAE    DYFKNSCP-AGYTWGRSFLFEDGAVCICNVDITVSVKE-NCIYHKSIFNGMNFPADGPVMKKMTTNWEASCEKIMPVPKQGILKGDV 165
1XA9    DYFKNSCP-AGYTWGRSFLFEDGAVCICNVDITVSVKE-NCIYHKSIFNGMNFPADGPVMKKMTTNWEASCEKIMPVPKQGILKGDV 165
1XSS    DYFKQSFP-EGYSWERSMSYEDGGICLATNNITMKKDGSNCFVYEIRFDGVNFPANGPVMQRKTVKWEPSTEKMYVRDG--VLKGDV 160
1ZUX    DYFKQSFP-KGYSWERSLTFEDGGICIARNDITMEG---DTFYNKVRFHGVNFPANGPVMQKKTLKWEPSTEKMYVRDG--VLTGDI 157
1BTJ    DYFKQSFP-KGYSWERSLTFEDGGICIARNDITMEG---DTFYNKVRFHGVNFPANGPVMQKKTLKWEPSTEKMYVRDG--VLTGDI 157
1GGX    DYKKLSFP-EGFKWERVMNFEDGGVVTVTQDSSLQD---GCFIYKVKFIGVNFPSDGPVMQKKTMGWEASTERLYPRDG--VLKGEI 161
1XQM    DYFKQSFP-EGFTWERTTTYEDGGFLTAHQDTSLDG---DCLVYKVKILGNNFPADGPVMQNKAERWEPATEILYEVDG--VLRGQS 158
2A54    DYFKQSFP-EGFTWERTTTYEDGGFLTAHQDTSLDG---DCLVYKVKILGNNFPADGPVMQNKAGRWEPSTEIVYEVDG--VLRGQS 158
1XMZ    DYFKQSFP-EGFTWERTTTYEDGGFLTAHQDTSLDG---DCLVYKVKILGNNFPADGPVMQNKAGRWEPGTEIVYEVDG--VLRGQS 158
1UIS    DFFKQSFP-EGFTWERVTRYEDGGVFTVMQDTSLED---GCLVYHAKVTGVNFPSNGAVMQKKTKGWEPNTEMLYPADG--GLRGYS 158
1MOU    DYVKQSFP-EGFTWERIMNFEDGAVCTVSNDSSIQG---NCFTYHVKFSGLNFPPNGPVMQKKTQGWEPHSERLFARGG--MLIGNN 157
            .    90    .  100    .  110    .     120    .  130    .  140    .  150    .    160

Monomeric
2H50    KMRLKLKDGGHYTSEVKTTYKA--KK--P-VQLPGAYIVGIKLDITSHNE-DYTIVEQYERAEGRHSTGGMDELYK- 236
Dimeric
1YZW    VMALKVGDRRLICHLYTSYRSK--KA-VRALTMPGFHFTDIRLQMPRKK--KDEYFELYEASVARYSDLPEKAN--- 227
2G6X    TRTFSLRDGGYYSSVVDSHMHFKSAIHPSILQNGGPMFAFRRVEE-DHS--N-TELGIVEYQHAFKTPD-------- 219
2G6Y    TRTFSLRDGGYYSSVVDSHMHFKSAIHPSILQNGGPMFAFRRVEE-DHS--N-TELGIVEYQHAFKTPD-------- 219
Tetrameric
2A48    TAFLMLQGGGNYRCQFHTSYKT--KK--P-VTMPPNHVVEHRIARTDLDK-GGNSVQLTEHAVAHITSVVPF----- 229
2A47    TAFLMLQGGGNYRCQFHTSYKT--KK--P-VTMPPNHVVETRIARTDLDK-GGNSVQLTEHAVAHITSVVPF----- 229
1ZGP    HKALKLKDGGHYLVEFKSIYMA--KK--P-VQLPGYYYVDSKLDITSHNE-DYTIVEQYERTEGRHHLFL------- 225
1XAE    SMYLLLKDGGRYRCQFDTVYKA--KSV-P-SKMPEWHFIQHKLLREDRSDAKNQKWQLTEHAIAFPSALA------- 231
1XA9    SMYLLLKDGGRYRCQFDTVYKA--KSV-P-SKMPEWHFIQHKLLREDRSDAKNQKWQLTEHAIAFPSALA------- 231
1XSS    NMALLLQGGGHYRCDFRTTYKA--KK--V-VQLPDYHFVDHRIEITSHDK-DYNKVKLYEHAKAHSGLPRLA----- 226
1ZUX    TMALLLEGNAHYRCDFRTTYKA--KE--KGVKLPGYHFVDHCIEILSHDK-DYNKVKLYEHAVAHSGLPD------- 222
1BTJ    TMALLLEGNAHYRCDFRTTYKA--KE--KGVKLPGYHFVDHCIEILSHDK-DYNKVKLYEHAVAHSGLPD------- 222
1GGX    HKALKLKDGGHYLVEFKSIYMA--KK--P-VQLPGYYYVDSKLDITSHNE-DYTIVEQYERTEGRHHLFL------- 225
1XQM    LMALKCPGGRHLTCHLHSTYRS--KKPASALKMPGFHFGDHRIEIMEEVE-KGKCYKQYEAAVARYCDAAPSKLGHH 232
2A54    LMALKCPGGRHLTCHLHTTYRS--KKPASALKMPGFHFEDHRIEIMEEVE-KGKCYKQYEAAVGRYCDAAPSKLGHN 232
1XMZ    LMALKCPGGRHLTCHLHTTYRS--KKPASALKMPGFHFEDHRIEIMEEVE-KGKCYKQYEAAVGRYCDAAPSKLGHN 232
1UIS    QMALNVDGGGYLSCSFETTYRS--KKTVENFKMPGFHFVDHRLERLEESD-KEMFVVQHEHAVAKFCDLPSKLGRL- 231
1MOU    FMALKLEGGGHYLCEFKTTYKA--KK--P-VKMPGYHYVDRKLDVTNHNK-DYTSVEQCEISIARKPVVA------- 221
           .  170    .  180      .     190    .  200    .   210    .  220    .
```

As shown in Table 5, different choices of the substitution matrix lead to similar performances, pointing to a good stability of the scheme. We believe that the (slightly) better performance of the Identity matrix is due to the presence in the set of several point mutants at some promising positions. This enrichment of local amino acid distribution brings a level of detail that can be better seized by a sharper discrimination between the different amino acids. Position 147 is assigned a star since it has been identified in experiments, but it can lead to non-fluorescent proteins [27, 30] and so it deserves a special classification. In Table 5 we also compared SDPhound results with those obtained by the SDPpred Web Server [2] using the same alignment and under the same conditions of shuffle number and of percentage of gaps exclusion. SDPpred returns a slightly lower number of hits and predicts only 9 of its positions to be statistically significant. Differences could arise from the different probability smoothing procedure and from the removal of background correlation, which does not seem to improve the results when performed on our set (see Sect. 4. of SI). The last row in the table reports how many of the 33 positions that are known to affect the aggregation state ranked higher than 33 in the various runs. As it can be seen from the Table, all runs succeed in pointing out the majority of the relevant positions with only minor fluctuations in the quality of the results.

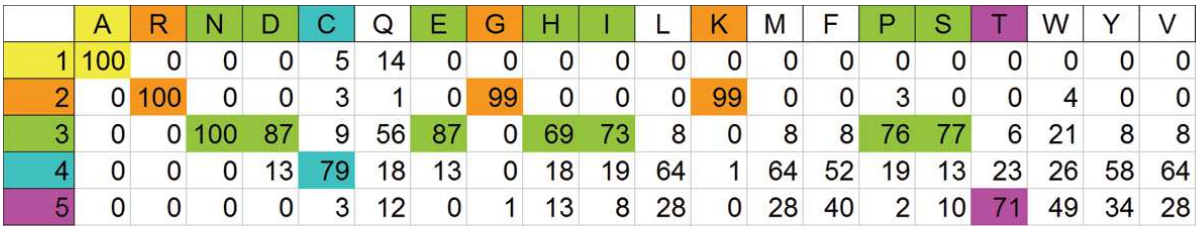
Table 6. Substitution class based runs over SPDs relevant for multimerization in IFPs family. Ranking and colors as in Table 2.

Pos.	**117**	83	**194**	**164**	**192**	156	223	175	153	5
hyd	4	–	1	8	2	3	22	7	15	–
crg	3	–	5	1	8	–	4	–	2	7
size	1	–	2	3	5	4	19	11	–	21
Pos.	**177**	**224**	**124**	**162**	**147**	**4**	**174**	**72**	**8**	**44**
hyd	–	11	–	–	34	–	–	–	–	6
crg	–	–	–	19	37	–	6	12	–	–
size	9	7	18	–	13	12	16	–	–	15

Looking for primary physical features

Having established a preliminary ranking of the putative SDPs, we address the question of identifying the relevant physical property/ies at each position, by means of the pigeonholing procedure devised above. Based on the results of the previous runs, and in order to give a more clear-cut meaning to the pigeonholes, we again chose the Identity matrix in the definition of the substitution probability, see eq. (14). In Table 6 we traced the 20 best ranking positions found in the Id column of Table 5 to investigate whether they were high ranked in one or more of those "per physical observable" based runs. Let us review the most significant results. Overall, pigeonholing in the IFP case once again picks many SPDs experimentally known to be relevant for the monomerization of IFPs; moreover, as will be clear in the following, such procedure once again offers valuable insight into the physical properties underlying the role of SDPs, with no other input but the sequence alignment. Hydrophobicity is important for positions 194, 192 and 156. It is apparent that nonpolar residues located at an interface can favor multimerization by decreasing the desolvation penalty of the external surface of the protein; this is the case for F194K or Y194K mutations. Charge is relevant for positions 164, 153, 117, and 223; in particular, mutations such as A164R or Y164R can introduce an electrostatic repulsion between monomers that prevents their aggregation. Steric hindrance can be a third cause for the decrease of the rate of binding between single units and the pigeonholing procedure spots positions 117, 194 and 164. It is worth noting that different physical features can be relevant for the same SDP, as is the case with positions 117, 194 and 164, thus restricting the possible identities of the corresponding residues. A164R mutation can, for instance, disrupt the hydrophobic packing of residues on the surfaces of monomers within a tetramer. In some cases, such as 83 and 8, none of the selected physical properties exhibits a significant score. This suggests that none of the partitionings is able to capture the peculiarity relevant in those last cases and it prompts to find alternatives to an a priori definition of the characteristics of the pigeonholes. With that in mind, the "optimal automatic pigeonholing" described in section 3.4. was used to generate new substitution classes. We applied it to the set of 92 sequences analyzed in subsection 5.2., setting the number of bins to 5, aiming to maximize the average mutual information of the 10 best ranking positions, regardless of which they are. The method provided a clear-cut partitioning only for some amino acids, as shown in Figure 3, more information on the run can be found in the caption. The obtained partitioning corresponds partially to the traditional amino acid groupings, such as the Taylor's Venn Diagrams [31]. In particular, ARG and LYS residues are grouped together and the individuality of CYS is revealed. Differences in partitionings, however, would not be new and have already been discussed [13]. Since we maximized the average information content of the 10 best ranking positions, the classes we sketch represent the intermingling of individual amino acid peculiarities that are relevant to the monomerization propensity of IFP; our procedure can also be utilized to focus on the properties of one single position at a time.

Figure 3. Results of the "automatic pigeonholing" procedure. The process reduces the alphabet from 20, ignoring the gaps, to 5 symbols (the rows in the table) while maximizing the average mutual information of the 10 best ranking positions. Interestingly, the corresponding reduction in the average information content is remarkably small: from 0.421 to 0.407. The space of possible partitionings was explored with 2.0 10^6 iterations. The table summarizes the results of the best 5% of amino acid partitionings: for each the percentage frequency of occupation for each class is shown. Colored residues have been unambiguously assigned, whereas the remaining ones show a more complex pattern of correlation.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	100	0	0	0	5	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	100	0	0	3	1	0	99	0	0	0	99	0	0	3	0	0	4	0	0
3	0	0	100	87	9	56	87	0	69	73	8	0	8	8	76	77	6	21	8	8
4	0	0	0	13	79	18	13	0	18	19	64	1	64	52	19	13	23	26	58	64
5	0	0	0	0	3	12	0	1	13	8	28	0	28	40	2	10	71	49	34	28

Table 7. SDPs inferred from 66 Di- and Tetrameric DsRed homologs. The Table shows the 24 best ranking SDPs. Scores discriminate experimentally validated positions(∗) from those responsible for generic multimerization(†) and those beneficial or even specific to the dimer-to-tetramer step(‡), according to [27]. Positions with different scores are colored differently to facilitate reading.

	B45	S	B62	S	Id	S	Bloc	S	SDPpred	S
	24		16		22		20		221	
1	124	*	124	*	127	‡	124	*	127	‡
2	127	‡	127	‡	153	*	172		124	*
3	156	*	172		125	‡	127	‡	172	
4	172		158		172		156	*	153	*
5	21	†	156	*	124	*	21	†	114	
6	158		21	†	114		114		217	†
7	114		6	†	6	†	180	†	125	‡
8	6	†	114		158		6	†	154	
9	44	†	181		156	*	160	†	156	*
10	225	*	211		154		158		76	
11	181		64		225	*	62		118	
12	146		87		176		170		167	
13	211		225	*	217	†	4		158	
14	64		44	†	209		211		21	†
15	87		146		39		36		87	
16	160		168		168		94		45	
17	192	*	180	†	45		202		39	
18	47		47		21	†	154		82	
19	170		192	*	208		175	*	155	
20	45		39		162	*	168		181	
21	107		4		163	†	192	*	62	
22	117	†	170		118		64		209	
23	147		160		180	†	209		163	†
24	4		202		167		225	*	176	

Pairwise correlation

As a final step in the analysis of the SDPs responsible for IFP monomerization, we investigated possible cooperative effects among different positions. This was done as described in 3.2. and using the Identity substitution matrix. The time required to perform runs for correlated effects with any BLOSUM-based similarity matrix is one order of magnitude larger than that needed when using the Identity. A pictorial representation of the symmetric correlation matrix is provided in the SI, a detailed listing of relevant pairs of positions id given in Figure 2 of SI; it is worth mentioning that an already known position, namely 177, as well as two "new" ones, 135 and 81, appear to consistently correlate with many other positions.

5.3. From tetrameric to dimeric form

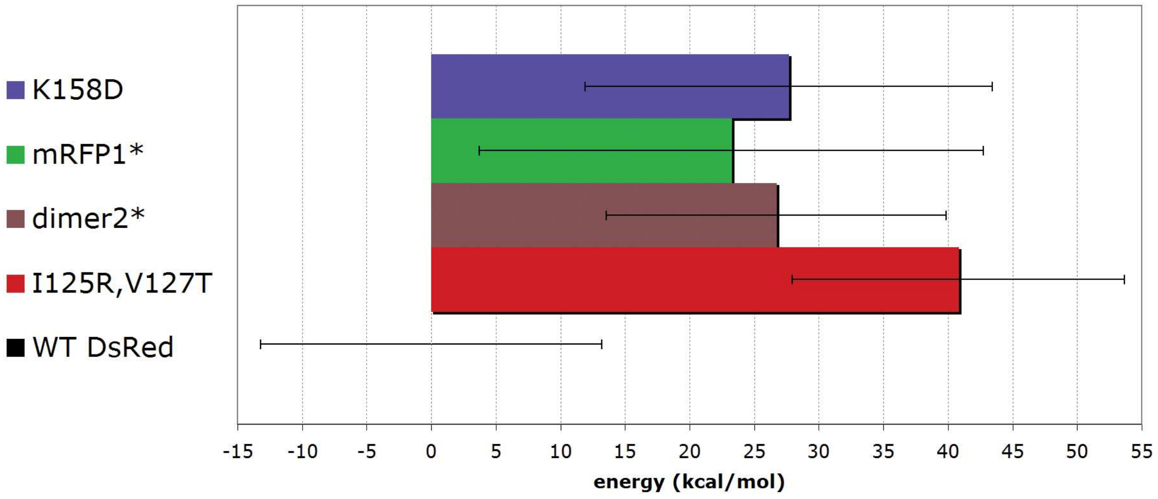
We now consider putative SDPs of the tetrameric vs. dimeric form of IFP and analyze some new mutations. To that end, two groups of 13 dimers and 53 tetramers, again sharing the same interface as in DsRed, (44.2% and 52.9% average identity, respectively) were selected from the alignment and underwent our procedure. Comparison between experiments and our results, reported in Table 7, is limited by the fact that the literature focuses on the generic distinction between monomers and multimers and that tetramer-to-monomer and tetramer-to-dimer roles very likely overlap. Out of the first 20 most significant positions predicted by the various flavours of the method implemented in SDPhound, including the BLOSUM based ones due to their potentially innovative character, we selected those not yet tried in experiments. Then, we investigated with the pigeonholing procedure (Table 8) the physical features influencing their specificity. We focused on position 158 since it ranks as the best non experimentally validated position in Table 7 that also belongs to the tetramerization interface in DsRed homologs. The results concerning SDP 158 in Table 8 indicate that dimerization is favoured by changes in size and, to a lesser extent, in charge. In particular, inspection of the sequences revealed that aspartic acid appears only in the dimeric class at position 158 (DTRKCI vs. TRKIV), leading us to select the K158D mutation as a promising candidate to further weaken dimer-dimer binding. We decided to obtain a further assessment of this mutation, validating it independently with the MMPBSA technique [10], a well-established and quite powerful tool to estimate binding free energies such as those involved in oligomerization processes. In this context, free energy is evaluated as an average over molecular dynamics trajectories of $G = E_{MM} + G_{PB,polar} + G_{SA,nonpolar} - TS_{solute}$.

Details concerning the meaning of the individual terms, the implementation and the actual calculation can be found in the SI. Here, we only point out that binding free energy ΔG_i can be estimated for each mutant as the free energy difference between tetramers and dimers, so that $\Delta\Delta G_{i,WT}$ is the binding free energy difference of the i^{th} mutant with respect to WT DsRed. To benchmark MMPBSA, we created, *in silico*, three mutants, namely I125R/V127T, dimer2* and mRFP1* (see caption of Figure 4 for details), for which the tetrameric character can be confidently excluded ($\Delta\Delta G_{i,WT} > 0$). Despite the rather short length (400 ps) of the dynamics (hence the relatively large error bars), the calculations, shown in Figure 4 together with the details about exact mutations, reproduce the expected trends and support the ability of the I125R mutation to disrupt the tetramer ($\Delta\Delta G_{I125R/V127T,WT}$=40.79 kcal/mol), consistently with what described in [27]. For dimer2* and mRFP1* we report slightly lower values ($\Delta\Delta G_{dimer2^*,WT}$=26.69 kcal/mol and $\Delta\Delta G_{mRFP1^*,WT}$=23.23 kcal/mol) than for the I125R/V127T mutant; this is reasonable since in those cases mutagenesis introduced many mutations that, in addition to providing control over oligomerization tendency, preserved other essential properties, such as correct and fast protein folding, chromophore maturation and detectable fluorescence. More interesting is the study of the tetramerization free energies of K158D mutant as well as the physical insights that emerge. $\Delta\Delta G_{K158D,WT}$=26.67 kcal/mol actually indicates a very promising mutation, not likely to affect the fluorescence due to the location of the residue far from the chromophore. Moreover, from the tetramer crystal one can see that there is a H-bond between the basic K158 of one monomer and the carbonyl group of N128 of the associating partner; such interaction only seems possible if position 158 hosts a large and positive residue, such as K or R, consistent with what indicated by Table 8.

Table 8. Substitution class based runs over SDPs relevant to tetramer to dimer transformation. Ranking and color coding as in Table 2. We chose to study the positions that had not been validated experimentally, since these are the new possibilities for putative mutations. For each position, its ranking in any specific run is shown. "–" means that in that run, the position ranked below 40th.

Pos.	172	158	114	181	211	154	64	62	146	87
hyd	1	–	–	16	–	–	32	6	14	–
crg	19	17	30	24	2	27	–	6	–	–
size	30	9	2	20	–	–	28	–	25	21
Pos.	**176**	**170**	**4**	**209**	**39**	**36**	**160**	**168**	**94**	**45**
hyd	29	–	31	25	–	21	–	–	–	–
crg	–	–	29	10	8	–	15	–	34	21
size	–	34	33	–	–	15	31	–	23	24

Figure 4. MMPBSA-estimated $\Delta\Delta G$s with respect to WT DsRed. Structures for all mutants were generated from crystal structure of this latter (1GGX PDB code). For dimer2* (T21S, H41T, C117T, I125R, V127T and S131P) and mRFP1*(T21S, H41T, C117E, I125R, V127T, R153E, V156A, H162K, A164R, L174D, I180T, Y192A, Y194K, H222S, L223T, F224G, L225A), only experimentally validated mutations according to [27] corresponding to solvent exposed residues were considered.



6. Conclusions

SDPhound, an articulate and flexible protocol for SDP identification and analysis within homologue proteins, is presented. Results of this work are twofold: $i)$ a methodological improvement over pre-existing techniques that use the mutual information to reveal SDPs, and $ii)$ the characterization of the role of known and unknown residues responsible for specificity. The procedure makes it possible to obtain useful insights from the data and to make sense of the results both in probabilistic and physical terms. It can be generalized to investigate correlations among position pairs. When applied to the standard test cases of the MPI and LacI families and to the original analysis of the multimerization tendency within the IFP family, SDPhound correctly identifies mutations that are recognized to be relevant and/or experimentally known to influence specificity. It further succeeds in pointing to the physical characteristics of several relevant residues in all the applications considered in this work. In the case of IFP it also suggests several previously uninvestigated positions as determinant for the multimerization state of these proteins, both in the case of the multimer to monomer transition and in that of the tetramer to dimer change. The hierarchy of operations described in sections 3. and 5. allows to draw a reliable and rather stringent profile of the set of residues that are responsible for the specific function, or structure, within a set of homologs, making SDPhound a useful tool to complement both experimental techniques and other computational biology approaches.

Acknowledgments

Funding from Italian MIUR (RBLA03ER38FIRB) and licensed MATLAB® access from Rome University "La Sapienza" are gratefully acknowledged. We thank A. Giansanti for interesting discussions on the method and careful reading of the manuscript, and K. A. Feenstra for the MIP and LacI alignments.

References and Notes

1. Pazos, F. and Bang, J.-W. Computational prediction of functionally important regions in proteins. *Curr. Bioinf.* **2006**, *1*, 15-23.
2. Kalinina, O.V.; Mironov, A.A; Gelfand, M.S; Rakhmaninova, A.B. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci* **2004**, *13*, 443-56.
3. Donald, J.E.; Shakhnovich, E.I. Predicting specificity residues in two large eukaryotic transcription factor families. *Nucl. Acids Res.* **2005**, *33*, 4455-65.
4. Bizzarri,R.; Nifosi, R.; Pingue, P.; Tozzini, V.; Beltram, F. Nano-Sized Optical "Devices" for Applications in Proteomics and Biomolecular Electronics: Engineered Green Fluorescent Proteins In *Functional Nanomaterials* Geckeler, K.E. and Rosenberg, E. Eds.; American Scientific Publisher: California, 2006; Chapter 2.
5. Chalfie, M.; Tu, Y.; Euskirchen, G.; Ward, W.W.; Prasher, D.C. Green fluorescent protein as a marker for gene expression. *Science* **1994**, *263*, 802-805.
6. Tozzini, V.; Pellegrini, v.; Beltram, F. *Handbook of organic photochemistry and photobiology*, Horsphool, W. M.; Lenci, F. , Eds.; CRC, Washington DC, 2004; Chapter 139.

7. Shimomura, O.; Johnson, F.H.; Saiga, Y. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *J. Cell. Comp. Physiol.* **1962**, *59*, 223-239.
8. Shagin, D.A; Barsova, E.V.; Yanushevich, Y.G.; Fradkov, A.F.; Lukyanov, K.A.; Labas, Y.A.; Semenova, T.N.; Ugalde, J.A.; Meyers, A.; Nunez, J.M.; Widder, E.A.; Lukyanov, S.A.; Matz, M.V. GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity. *Mol. Biol. Evol.* **2004**, *21*, 841-850.
9. Shaner, N.C.; Steinbach, P.A.; Tsien, R.Y. A guide to choosing fluorescent proteins. *Nature Methods* **2005**, *2*, 905-909.
10. Kollman, P.A.; Massova, I.;, Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D.A.; Cheatham, 3rd., T.E. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889-897.
11. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *96*, 10915-10919.
12. Mirny, L.A.; Gelfand, M.S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* **2004**, *321*, 7-20.
13. Thompson, M.J., Goldstein, R.A. Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes. *Prot. Struct. Funct. Gen.* **1996**, *25*, 38-47.
14. Tozzini, V. Coarse Graine Models for Proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144-150
15. Henikoff, S.; Henikoff, J.G. Position-based sequence weights. *J. Mol. Biol.* **1994**, *243*, 574578.
16. Tsujishita, T. On triple mutual information. *Advances in applied mathematics* **1995**, *16*, 269-274.
17. McGill, W.J. Multivariate Information Transmission. *IEEE Trans. Information Theory* **1954**, *4*, 93-111.
18. Good, P. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer Series in Statistics, Springer: New York, 1994.
19. Mirny, L.A.; Gelfand, M.S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J. Mol. Biol.* **2002**, *321*, 7-20.
20. Pirovano, W.; Feenstra, K.A.; Heringa, J. Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res.* **2006**, *34*, 6540-6548.
21. Kalinina, O.V.; Novichkov, P.S.; Mironov. A.A.; Gelfand, M.S.; Rakhmaninova, A.B. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acid Research* **2004**, *32*, W424-W428.
22. Fu, D.; Libson, A.; Miercke, L.J.; Weitzman, C.; Nollert, P.; Krucinski, J.; Stroud, R.M. Structure of a glycerol-conducting channel and the basis for its selectivity. *Science* **2000**, *290*, 481-486.
23. Sui, H.; Han, B.G.; Lee, J.K.; Walian, P.; Jap, B.K. Structural basis of water-specific transport through the AQP1 water channel. *Nature* **2001**, *414*, 872-878.
24. Lu, F.; Schumacher, M.A.; Arvidson, D.N.; Haldimann, A.; Wanner, B.L.; Zalkin, H.; Brennan, R.G. Structure-Based Redesign of Corepressor Specificity of the Escherichia coli Purine Repressor by Substitution of Residue 190. *Biochem.* **1998**, *37*, 971-982.

25. Glasfeld, A.; Koehler, A.N.; Schumacher, M.A.; Brennan, R.G. 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*, 347–361.
26. Schumacher, M.A.; Choi, K.Y.; Zalkin, H.; Brennan, R.G. Crystal structure of LacI member PurR, bound to DNA: minor groove binding by alpha helices. *Science* **1994**, *266*, 763-770.
27. Campbell, R.E.; Tour, O.; Palmer, A.E.; Steinbach, P.A.; Baird, G.S.; Zacharias, D.A.; Tsien, R.Y. A monomeric fluorescent protein. *Proc. Natl. Acad. Sci. USA.* **2002**, *99*, 7877-7882.
28. Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Molec. Graphics* **1996**, *14*, 33-38.
29. Russell, R.B.; Barton, G.J. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* **1992**, *14*, 309-323.
30. Baird, G.S. PhD Thesis (University of California, San Diego) *2001*.
31. Taylor, W. The classification of amino acid conservation. *J. Theor. Biol.* **1986**, *119*, 205-218.