PAPER

CrossMark
click for updates

Identifying relevant positions in proteins by Critical Variable Selection†

Silvia Grigolon,[a] Silvio Franz[b] and Matteo Marsili*[c]

Evolution in its course has found a variety of solutions to the same optimisation problem. The advent of high-throughput genomic sequencing has made available extensive data from which, in principle, one can infer the underlying structure on which biological functions rely. In this paper, we present a new method aimed at the extraction of sites encoding structural and functional properties from a set of protein primary sequences, namely a multiple sequence alignment. The method, called critical variable selection, is based on the idea that subsets of relevant sites correspond to subsequences that occur with a particularly broad frequency distribution in the dataset. By applying this algorithm to *in silico* sequences, to the response regulator receiver and to the voltage sensor domain of ion channels, we show that this procedure recovers not only the information encoded in single site statistics and pairwise correlations but also captures dependencies going beyond pairwise correlations. The method proposed here is complementary to statistical coupling analysis, in that the most relevant sites predicted by the two methods differ markedly. We find robust and consistent results for datasets as small as few hundred sequences that reveal a hidden hierarchy of sites that are consistent with the present knowledge on biologically relevant sites and evolutionary dynamics. This suggests that critical variable selection is capable of identifying a core of sites encoding functional and structural information in a multiple sequence alignment.

1 Introduction

The structure and function that proteins perform inside cells is encoded in their amino acid sequence.[1] Yet sequences are subject to biological evolution, hence the same protein or protein domain may correspond to remarkably different sequences for organisms that are far apart in the evolutionary tree. What constrains evolution is precisely the requirement that the structure and biological function be conserved.[2–4] The sequence of a given protein in different organisms can be regarded as a collection of "solutions" that evolution has found to the same optimisation problem. This observation lies at the basis of methods for inferring the way in which the structure and functions are encoded in the sequences of amino acids across different species.[5] The first step consists in compiling a database of sequences for a given protein (domain) across species, called Multiple Sequence

Alignment (MSA).‡ Secondly, decoding the statistical traces that constrained evolution leaves in the MSA allows one to reverse engineering those positions that play a relevant role. Thus, the frequency of mutations on single sites reveals those positions along the sequence that are "protected" from mutations, either because they are associated with important biological functions or because they are vital for the stability of the tertiary structure. Furthermore, correlations in the mutation of pairs of sites carry information that can be reverse engineered to reveal contacts in the 3D structure.[5–12] Coevolution on larger subsets of sites can be spotted by an extension of principal component analysis, called Statistical Coupling Analysis (SCA)[13], that aims at identifying regions that are associated with functional domains.[8,13]

Yet, in all these examples, inference techniques are limited by and rooted in statistics that do not go beyond pairwise correlations. Indeed, available data barely allow one to estimate pairwise correlations, not to speak of higher order statistics. Yet, there is no reason why evolution should use only pairwise correlations to encode biological functions in amino acid sequences. As a matter of fact, selection operates on the whole sequence.

[a] *The Francis Crick Institute, Lincoln's Inn Fields Laboratory, 44, Lincoln's Inn Fields, London WC2A 3LY, UK. E-mail: silvia.grigolon@gmail.com*

[b] *LPTMS – Laboratoire de Physique Théorique et Modèles Statistiques, UMR CNRS 8626, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay, France. E-mail: silvio.franz@lptms.u-psud.fr*

[c] *Abdus Salam International Center of Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy. E-mail: marsili@ictp.it*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6mb00047a

‡ Alignment is necessary in order to compare different positions along the sequence. Here we take the alignment to be given, to avoid entering into the intricacies of algorithms for multiple sequence alignment.

In this paper, we propose a new statistical non-parametric method going beyond pairwise correlations for the analysis of MSAs of a given protein (domain) family. The method, that we call critical variable selection (CVS), is based on the conclusions of ref. 14 that sampling of relevant degrees of freedom of a complex system generates datasets with broad distribution of frequencies.§ This conclusion is further refined in ref. 15 within a Bayesian model selection approach. This paper is the first systematic attempt to exploit these ideas for identifying the subset of relevant positions in a biological dataset. Unlike other methods, CVS aims at distinguishing relevant from irrelevant variables, purely on information theoretic grounds, without assuming an *a priori* criterion of relevance. The aim of our method is to unveil first of all whether there is a well-defined hierarchy in a given set of variables and, as a consequence, to characterise it. Here, relevance becomes a sample specific concept, depending on the way the sample has been assembled. As an example, if no hierarchy is present among the variables, we would not expect CVS to identify a distinction between relevant and irrelevant. In this sense, our method differs from methods such as Direct Coupling Analysis (DCA)[7] aimed at identifying specific features (*e.g.*, contacts between amino acids), that call for models (with pairwise interactions) in which those features are related to sufficient statistics. The downside of our approach is that, while it is easy to validate methods that target a specific goal for hundreds of protein families (*e.g.*, by comparing DCA predictions with measured inter-residues distances), the validation of our method requires in-depth analysis of the protein family, because what is relevant in one family needs not be relevant in another. Therefore, our analysis will focus on two specific families, the Response Regulator Receiver (RR, PFAM ID PF00072) and the voltage sensor domain of the ion channels (VSD, PFAM ID PF00520), though it has been performed on several other families. The sequences in these two families have nearly the same length, but the size of the database differs by an order of magnitude. This allows us to probe CVS as the depth of the dataset varies, which is an important dimension.

The paper is structured as follows. After a general introduction into the method's formulation and the related algorithm, we show the outcome of CVS when applied to an *in silico* sequence, a paradigmatic example to understand what kind of results such method affords. We then proceed with the study of real biological sequences, showing that CVS can consistently identify the existence of an underlying ranking in relevance among sites on actual samples. These preliminary sections are then followed by more in depth analyses, showing the ability of CVS to unveil information going beyond pairwise correlations and the robustness of the method with respect to sampling and evolutionary biases. We then compare our results with the predictions of state-of-the-art methods, such as single-site conservation, Statistical Coupling Analysis (SCA)[13] and Direct Coupling Analysis (DCA).[5–7] Our whole analysis shows that, in the cases studied, CVS identifies a core of interdependent positions that are denser and tighter than those identified by SCA. Finally, we discuss the biological relevance of the sites identified by CVS in the response regulator receivers and the voltage sensor domain of the ion channels, in relation to site conservation, functional sites and solvent accessible surface scores.

2 Critical variable selection

2.1 Theory

Let us consider multiple sequence alignment of homologous proteins composed of M sequences of length L, $\vec{s}^{\alpha} = (a_1^{\alpha}, \ldots, a_L^{\alpha})$, where $\alpha = 1, \ldots, M$ labels the sequence and a_i^{α} is the ith amino acid of the αth sequence.¶ Each sequence $\vec{s} = (a_1, \ldots, a_L)$ can be thought of as a "solution" of how the same biological function is achieved in the specific environment where that particular sequence has evolved. Following ref. 14, we think of this as an optimisation problem of an unknown function that also depends on unobserved variables. Ref. 14 suggests that the frequency with which a given sub-sequence occurs provides information on the function being optimised. Highly conserved sites are expected to be functionally relevant, whereas sites with high variability are unlikely to be relevant. While the former correspond to a peaked frequency distribution, where few sub-sequences occur very often, a subset of highly variables sites yield a flat distribution of frequencies with many sub-sequences occurring a few times. So, for a given sub-set of sites, the variability in the frequency distribution is a proxy of the variability on those sites of the function being optimised. Therefore, the most informative sub-sequences are those for which the frequency distribution has the broader variation. This suggests looking for subsets of variables such that the frequency with which the corresponding subsequences occur has a larger variability in the MSA. This idea is best illustrated with a specific example, referring interested readers to ref. 14 and 15 for a more detailed theoretical discussion.

Let the MSA in Fig. 1(a) be a compilation of sequences across the evolutionary tree for protein domains that perform the same biological function. Notice that each sequence $\vec{s}^{\alpha}$ is unique, which reflects the fact that the MSA should sample as uniformly as possible the evolutionary process. Let us assume that, of the $L = 12$ positions, $n = 6$ are relevant for the function of interest and the remaining ones are irrelevant. Fig. 1(b and c) report two subsets $I_b = \{2, 3, 9, 10, 11, 12\}$ and $I_c = \{2, 3, 4, 9, 11, 12\}$ of positions. For each $I \subseteq \{1, 2, \ldots, L\}$ we split the sequence $\vec{s} = (\underline{s}_I, \bar{s}_I)$ in the subsequence $\underline{s}_I = \{a_i : i \in I\}$ over the (putatively) relevant sites and in the subsequence $\bar{s}_I$ of the remaining sites. Let

$$k_I(\underline{s}) = \sum_{\alpha=1}^{M} \delta_{\underline{s}, s_I^{\alpha}}$$

¶ The values that a_i^{α} can take correspond to the different amino acids. Including gaps and ambiguous amino acids, a_i^{α} can take up to 21 values in the examples studied in this paper.
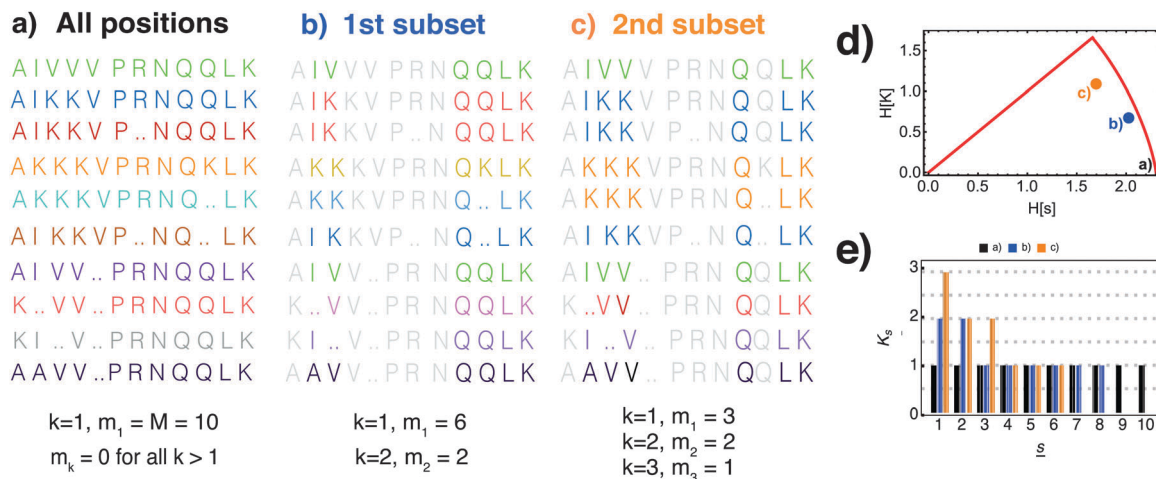
Fig. 1 (a)–(c) Example of three samples obtained from the same one when considering different positions, highlighting the corresponding different states, $\underline{s}$. Counts k and their multiplicities m_k are shown as well below the three panels. (d) Relevance $H[K]$ as a function of the resolution, $H[\vec{s}]$, for the three samples represented in (a)–(c) below the red line depicting the Poissonian case. (e) Distributions of the counts, $K\vec{s}$, for the different states $\underline{s}$ observed in the three samples in (a)–(c).

be the frequency number of times with which different sub-sequences $\underline{s}_I$ occur in the sample. As Fig. 1 shows, different choices of I lead to different distributions of frequencies, that in general differ from the flat frequency distribution $k(\underline{s}) = 1$ corresponding to the entire sequence $\vec{s}$ (black histogram in Fig. 1(e)). The key observation in ref. 14 is that a broader distribution of frequency (the one corresponding to I_c here) will be more informative of the function being optimised, because it allows us to distinguish more states in terms of their "functionality". A quantitative measure of relevance is then given by the entropy of the frequency distribution

$$H[K_I] = -\sum_k \frac{km_I(k)}{M}\log\frac{km_I(k)}{M} \qquad (1)$$

where

$$m_I(k) = \sum_{\underline{s}} \delta_{k,k_I(\underline{s})} \qquad (2)$$

is the number of subsequences $\underline{s}_I$ that occur k times in the MSA. $H[K_I]$ is a proxy for the (log of the) number of states that the subset I allows one to resolve in frequency. Notice that $H[K_I]$ is different from the entropy of the sequence $\underline{s}_I$

$$H[\underline{s}_I] = -\sum_{\underline{s}} \frac{k_I(\underline{s})}{M}\log\frac{k_I(\underline{s})}{M} = -\sum_k \frac{km_I(k)}{M}\log\frac{k}{M}. \qquad (3)$$

While $H[\underline{s}_I]$ measures resolution, $H[K_I]$ measures relevance. Ref. 15, to which we refer, gives further compelling arguments for this conclusion on the basis of a Bayesian model selection framework, showing that $H[K]$ correlates with the number of parameters that can be estimated on the basis of the data. In this respect, finding the subset I that has maximal $H[K_I]$ is tantamount to finding those variables that are described by the richest model.

Going back to our example, Fig. 1(d) shows the values of $H[\underline{s}]$ and $H[K]$ for the frequency distributions (a)–(c) in the figure (we will suppress the index I if that creates no ambiguity

in what follows). The solid line provides an upper boundary of the relation between $H[\underline{s}]$ and $H[K]$.[14] Notice that the inclusion of totally conserved sites, such as {9, 11, 12} does not affect the value of either $H[\underline{s}]$ and $H[K]$, whereas inclusion of position 4, that covaries with 2 and 3, engenders a smaller decrease in $H[K]$ with respect to position 10.

In general, we shall posit that the relevance of a sub-set of positions I is quantified by $H[K_I]$. The aim is thus to find the sub-set I of not necessarily contiguous positions that maximises $H[K_I]$. As in the example mentioned above, we shall look for maxima over the sub-set I of $n = |I|$ positions:

$$I_n^* = \arg\max_{I:|I|=n} H[K_I]. \qquad (4)$$

The prediction of relevant sites based on the maximisation of $H[K]$ will be called Critical Variable Selection (CVS) in what follows. CVS aims at finding sub-sets of sites with a high degree of interdependency, i.e. with correlations extending to the whole dataset. Such large scale correlations are usually associated with power law frequency distributions. Indeed, ref. 14 shows that frequency distributions that maximise $H[K_I]$ at a given resolution $H[\underline{s}_I] = \bar{H}$, typically exhibit a power law frequency distribution $m_I(k) \propto k^{-\mu-1}$. We note in passing, that the distribution that maximises $H[K]$ differs from widely studied sampling distributions, such as the Ewens formula.[16] The algorithm that searches for solutions of this problem is described below. Before then, it is worth remarking on a few important points.

First, relevance is a relative concept. Here it is relative to the criterium with which the sequences of the MSA have been selected. The quality of the results depends ultimately on the quality of the data and on the algorithms for the multiple sequence alignment that is used. This means that one should not expect a structure to be unveiled from data if it is not present a priori in the data. We shall not discuss issues related to MSA algorithms and rely on MSA curated and compiled by others[8,17] that we regard as benchmarks.

At odds with other methods,[6,13] CVS does not rely on an explicit definition of correlation between sites. Indeed, it is based on statistics (the frequency) that goes beyond pairwise correlation and it uncovers information that goes beyond pairwise correlations.

Furthermore, this method assumes as the only parameter, the number n of positions that sets the resolution in the optimisation problem. Changing n allows us to see how the set of relevant sites expands. If the MSA contains a hidden hierarchy of relevant sites, we expect that varying n is equivalent to "zooming" in and out of the subset of relevant sites, thereby revealing the hierarchy. We also expect this hierarchy to be corrupted by noise, when n becomes too large. So the parameter n will be used for a mere aid in the search for a hierarchy of relevant sites.

2.2 The algorithm

Following our previous considerations, in this section we want to propose an operative way of identifying the subsets of n positions, $I^*(n)$, maximising the relevance function $H[K]$. In order to implement eqn (4), we employ a simple greedy gradient ascent algorithm: starting from a random choice of the subset I of n positions, we iterate the following steps:

(1) Construct a new tentative subset I' by picking one position i at random in the subset I and exchange it randomly with another position i' that is not in I;

(2) compute $H[K_{I'}]$ and $H[K_I]$;

(3) if $H[K_{I'}] \geq H[K_I]$, accept the move to the new position in the subset, i.e., $I \to I'$, otherwise leave I unchanged.

The algorithm stops when $H[K_I]$ does not change for a sufficiently large number of (attempted) moves.‖ Typically, for a given value of the subsequence length n, the algorithm does not produce a single maximum but rather a population of local maxima with similar values of $H[K]$, calling for the need for appropriately sampling the different possible solutions. In order to get a consistent statistics and fully explore such maxima profiles, we run the algorithm R times for each value of n starting from randomly chosen subsets I (typically $R \simeq$ 100–1000). Therefore the algorithm returns a distribution of solutions, each corresponding to a local maximum of $H[K_I]$. In order to assess the relevance of each position, we count the number $c_i(n)$ of times that position i is selected in the R solutions. We shall call this value simply the count. Upon running the algorithm for different values of n, each position can be assigned a total count, given by the sum of the single counts obtained for each n, i.e., $C_i = \sum_n c_i(n)$, which provides information on the "relevance" landscape of a given protein family.

3 The data

In order to understand the typical output of our method, we applied it to three different datasets, namely one *in silico* family of sequences and two biological MSAs. The former is made of

‖ For the examples studied in this paper, $20L$ steps were typically found to be enough for these conditions to be met.

$M = 10^4$ sequences of length $L = 64$ where, for simplicity, each amino acid a_i^α, $i = 1,\ldots,L$, $\alpha = 1,\ldots,M$ can take only two different values, i.e., either 0 or 1. Each sequence in the dataset is generated in order to include sites with a different degree of conservation and of mutual dependence. Each generated sequence is divided into four regions:

• a core made of 5 highly correlated positions a_i^α, $i = 1,\ldots,5$. In order to create non-trivial correlations that go beyond second-order statistics, these values are assigned taking the first five bits in the binary representation of a random number $X \in [0, 1]$ drawn from the pdf $p(x) = 1/(2\sqrt{x})$;

• a set of 12 subordinated sites that take values that are noisy functions of the core variables. For $i = 6,\ldots,15$, defining $s_i^\alpha = 2a_i^\alpha - 1$,

$$P\{s_i^\alpha = s_k^\alpha s_l^\alpha s_n^\alpha\} = 0.95,$$

where $1 \leq k < l < n \leq 5$ take all possible combinations, whereas s_{16}^α and s_{17}^α take values $s_{16}^\alpha = s_2^\alpha s_3^\alpha s_4^\alpha$ and $s_{17}^\alpha = s_2^\alpha s_3^\alpha s_4^\alpha s_5^\alpha$ in 95% of the cases and the opposite values otherwise.

• a set of highly conserved sites made of 17 positions, where s_i^α takes the same value in 95% of the cases.

• a set of random sites made of the remaining 31 positions, obtained by drawing each of them, independently at random with $P\{s_i^\alpha = 0\} = 0.5$.

An example of this dataset is shown in Fig. 2(a).

The biological MSAs analysed in the following refer instead to two different protein families, i.e., the response regulator receiver and the voltage sensor domain of the ion channels.

Response regulator receivers (RR, PFAM ID PF00072) are part of two-component signal transduction machineries allowing cells to sense and respond to a high variety of environments. These two components are usually made of a histidine kinase (HK) aimed at sensing the surrounding environment, i.e., controlling the input, whose signal is received by the response receiver domain (RR), that in turn triggers cell physiology and response.[18] The RR consists of two other subdomains, a N-terminal response regulator receiver domain and a variable C-term domain aimed at DNA binding. RRs adapted to a wide variety of signals and their evolutionary spread as well as their availability make this system suitable for any statistical analysis, and in particular for our technical purposes. Within the context of our analysis, we focused on the dataset used in ref. 8, made of $N = 62\,074$ sequences of length $L = 112$.

Voltage-dependent ion channels are biomolecular machines aimed at measuring changes in the cell transmembrane voltage and, because of their ubiquity, are found to be as well connected to many heritable diseases.[19] Structurally, these channels are made of four identical subunits, the voltage sensors, each of them divided in turn into six segments (S1–S6). Hereby, we are going to focus on the first four segments building up the so-called Voltage-Sensor Domain (VSD), (S1–S4), found to undergo conformational changes during voltage sensing. Our dataset is made of $M = 6652$ sequences of length $L = 114$, already curated in ref. 17. Note that the size of this dataset is much lower than that of the response regulator receiver, although the length of
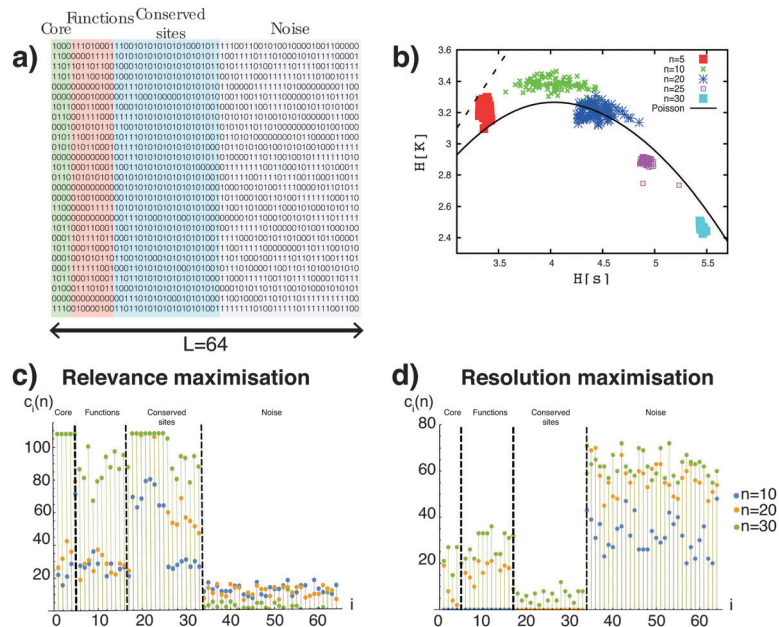
Fig. 2 (a) Example of a typical sample of the *in silico* sequences used to test critical variable selection. The different parts constituting each sequence are highlighted as described in the text. (b) Relevance, $H[K]$, as a function of the resolution, $H[s]$, for the sites maximising the relevance varying the subsequence length n. The results for Poissonian-distributed multiplicities m_k are shown as well.[14] (c) Single site counts for different subsequence lengths, $c_i(n)$, obtained by maximising the relevance, $H[K]$. The results obtained here are highly different than those shown in (d) as the most frequent selected positions are indeed the ones belonging to the core and the conserved sites. (d) Single site counts for different subsequence lengths, $c_i(n)$, obtained by maximising the resolution, $H[s]$. As described in the text, such a quantity tends to select by defining those positions showing maximal variability across the dataset.

the sequences is similar. This sample can be then informative for testing the ability of our method to extract relevant information even in the strong undersampling regime.

4 Results and discussion

4.1 Typical behaviour of the algorithm

In this section we illustrate the behaviour of the CVS algorithm, by applying it to synthetic and biological MSA.

4.1.1 *In silico* sequences. Let us now focus on the typical behaviour and the outcomes of this algorithm when run on the *in silico* sequence dataset built *ad hoc* to understand the typical output of critical variable selection. Fig. 2(a) shows a sample of the dataset. As described before, besides highly conserved sites, the synthetic MSA also contains sites with non-trivial correlation, going beyond second order statistics. We run the CVS algorithm 100 times for n = 5, 10, 15, 20, 25 and 30. Fig. 2(b) shows the values of $H[s]$ and $H[K]$ obtained in each run, for each value of the sub-sequence length n. With increasing n, CVS returns subsets of sites with increasingly higher resolution $H[\underline{s}]$. At the same time, $H[K]$ exhibits a non-monotonic behaviour as it first increases from n = 5 to n = 10 and then decreases. Close to the maximum, the points corresponding to different runs are scattered over a large region, suggesting that CVS prediction is the most noisy. Indeed, the optimal value of n does not correspond to the maximum of $H[K]$ because n is too small to adequately capture the statistical dependences in the MSA. Much insight is provided by looking at how the samples of

optimal sub-sets of positions evolve as n varies. Fig. 2(c) shows the counts $c_i(n)$ for each position, for n = 10, 20 and 30. CVS first identifies highly conserved sites, together with those core sites that are most conserved. Next, core sites and the subordinated sites are selected when n increases. Notice that the separation between relevant and irrelevant sites becomes sharper and sharper as n increases: while the counts of relevant sites increase with n, those of random sites decrease as n increases. This fact implies that sites that are found relevant by CVS at a given n remain relevant when n increases. This is a strong indication that CVS is uncovering a hidden hierarchy of relevant positions. This hierarchy is revealed by the ranking of positions in terms of total counts C_i. The top ranked sites turn out to be the highly conserved ones, followed by the core sites and the functional sites. In this case, the noisy part is very poorly ranked.

Fig. 2(d) contrasts these results with what one obtains by maximising the resolution $H[\underline{s}]$ instead. This procedure is expected to select highly variable sites. Indeed, we observe the opposite scenario, where counts $c_i(n)$ increase with n for the random sites and decrease with n for conserved and functional sites.

In summary, as n increases, CVS reveals the hierarchy of conservation and dependence that is hidden in the MSA. In order to reveal this structure it is important to run CVS for increasing values of n, at least as long as the separation between relevant and irrelevant sites remains sharp.

4.1.2 Biological sequences: sample size, reshuffling and reweighing. While in the study of *in silico* MSA the ground truth of the generating algorithm is available for the validation of results, for real biological sequences one has to resort to

information provided by different approaches, such as annotations of known functional sites or structural properties.

Before doing that, it is worth discussing the performance of CVS with respect to the following important aspects of the MSA: (i) sample size M, (ii) evolutionary bias and (iii) the nature of correlations.

We illustrate these aspects for the MSA of the RR domain. The analysis presented here can be considered as a preliminary study that can be applied to any MSA in order to assess the statistical robustness of the output of CVS.

The emergence of a hierarchy of relevant sites can be spotted by ranking sites according to their total count C_i and plotting the counts $c_i(n)$ as a function of their rank, for different values of n, as shown in Fig. 3(a). As a quantitative measure of the sharpness of the separation between relevant and irrelevant sites, we calculate dispersion, which is defined as

$$\text{Dispersion} = \frac{1}{L}\sum_i 4p_i(1-p_i) \qquad (5)$$

where $p_i = \dfrac{c_i}{R}$ is the probability of site i to be selected by CVS, estimated as the fraction of times position i is selected in the R independent runs. A small value of the dispersion implies that all sites are either selected in most runs ($p_i \simeq 1$) or rarely selected ($p_i \simeq 0$). As a benchmark, in the random case we expect counts to be $c_i \simeq n/L$ so that when CVS is purely dominated by noise the dispersion is given by $\approx 4(n/L)(1 - n/L)$.

We use this measure to test the robustness of CVS with respect to the sample size M. Starting from the original dataset for RR with poorly gapped sequences ($M_0 = 47\,349$ and the

number of gaps for each sequence is less than 4%) we iteratively reduce M by a factor of two, by randomly selecting a subset of half of the sequences each time. Fig. 3(b) shows the behaviour of the dispersion, for $n = 40$, as a function of the sample size M. For $M \approx 100$ this approaches the random limit ≈ 0.92, whereas for $M > 600$ a well defined structure emerges, which becomes very robust for $M > 2000$.

A second useful measure is the overlap

$$\text{Overlap} = \frac{1}{L}\sum_i \left[p_i^0 p_i^1 + \left(1 - p_i^0\right)\left(1 - p_i^1\right)\right], \qquad (6)$$

between the outputs p_i^0 and p_i^1 of CVS for two different datasets. This has the simple interpretation as the probability, averaged over sites, that CVS predicts the same relevance for site i in the two datasets. In Fig. 3(b) p_i^0 refers to CVS for the full dataset whereas p_i^1 refers to the reduced dataset with M sequences. In the random case, we expect an overlap ≈ 0.54, which is the value we observe for $M \simeq 100$. For larger values of M we see that the overlap converges to values close to one for $M > 2000$. In summary, for $M > 2000$ we observe a well defined structure of relevant sites, and this structure is the same up to $M = 47\,349$.

A second aspect that requires particular attention is the bias in the datasets which comes from the fact that sequences belonging to better studied organisms typically occur with enhanced frequencies. In order to partially correct for this bias, we limit our analysis to MSAs of sequences that differ in more than d positions.** Fig. 3(c) shows the dispersion for different realisations of the dataset as a function of the length of the subsequence, n, whereas Fig. 3(d) shows a density plot of the
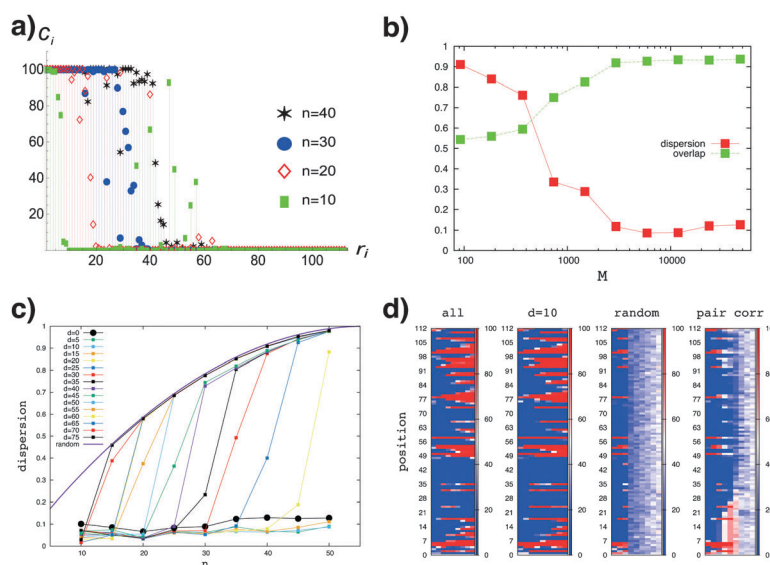

Fig. 3 (a) Critical variable selection outcome for the response regulator receiver dataset: single-site counts, $c_i(n)$ as a function of their rank r_i with respect to single-site total counts C_i, for $n = 10, 20, 30$ and 40 ($R = 100$). Different colours refer to different subsequences' lengths as shown in the side legend. (b) Overlap and dispersion of the CVS outcome on datasets built up taking into account M sequences of the actual one (legend is shown as the inset). (c) Dispersion between CVS outcome on different realisations of the dataset and the randomly picked sites as a function of the subsequence length n. Here the different realisations correspond to the actual dataset collapsed by using a similarity threshold equal to d amino acids (the legend is shown as the inset). (d) Heat maps showing the single-site counts, $c_i(n)$, as a function of the subsequence length $n = 10, 15, \ldots, 50$ (x-axis) and the position along the sequence (y-axis) for four different realisations of the dataset, $i.e.$, the actual one, the one where sequences are collapsed by a similarity threshold $d = 10$ amino acids, the reshuffled dataset constraining single-site frequencies and the reshuffled dataset constraining pairwise correlations.

counts $c_i(n)$ as a function of the position i and of n, for the cases $d = 0$ and $d = 10$ (first two panels). The separation between relevant and irrelevant sites predicted by CVS turns out to be sharper for $d = 5$, 10 and 15 than in the original dataset (smaller dispersion). This confirms that biases in the sampling of the evolutionary process indeed mask statistical dependences. For larger values of d, important traces of the evolutionary dynamics are lost, as suggested by the fact that the dispersion for large n converges to the random limit (*e.g.* for $d = 30$ and $n \geq 40$). This analysis suggests that $d = 10$ is a meaningful threshold for the similarity of sequences.

Finally, in order to gauge the nature of the statistical dependencies between positions, we compare the predictions of CVS in the original dataset to those obtained from randomly reshuffled MSAs.

First, in order to understand how much single site conservation affects the results of CVS, we produced a random MSA where, at each position, the amino acids were reshuffled across sequences. Secondly, in order to test the relevance of pairwise correlations, we followed the simulated annealing procedure of ref. 20 that allows us to produce randomised datasets that preserve pairwise correlations.

The heat maps of the counts $c_i(n)$ in Fig. 3(d) show that on the datasets randomised by constraining single site frequencies (third panel) and correlations (fourth panel) CVS is able to infer a hierarchy if $n \leq 15$ and $n \leq 30$, respectively. This translates into the fact that for $n \leq 15$ the structure within the dataset is mainly dominated by single site conservation and then by pairwise correlations up to $n \simeq 30$, generating predictions which are close to random (*i.e.* $c_i(n) \approx Rn/L$) for $n > 30$. By contrast, CVS maintains a low dispersion on the original MSA and on the $d = 10$ MSA up to $n = 50$ and beyond, highlighting the presence of statistical dependencies in the evolutionary process going well beyond pairwise correlations.

Finally, it is also worth noticing that single site conservation does not necessarily imply relevance. Site 112 in RR formally appears to be highly conserved in the dataset because it is very often a gap. Interestingly, while this site is picked up as relevant in the reshuffled MSAs, its counts are negligible for $n > 20$ in the real dataset.

In summary, this analysis allows us to draw three main conclusions on this specific MSA: (i) critical variable selection provides a robust choice of sites over a wide range of sample sizes; (ii) critical variable selection provides robust results and selects non-random sites when reweighting sequences with a similarity threshold up to $d = 15$ amino acids; (iii) critical variable selection infers a signal in the dataset going beyond pure single-site conservation or pairwise correlations.

This analysis can be performed on any MSA corresponding to a given protein domain family, and it provides a preliminary

test on whether CVS extracts a robust and non-trivial information from the MSA. The next step is to analyse the biological relevance of the results. But before doing that, we will compare and combine CVS to other methods for the analysis of MSAs.

4.2 Comparison with different statistical methods: statistical coupling analysis and direct coupling analysis

To better understand the behaviour of CVS, we compare its predictions to those of well-established statistical techniques aimed at identifying relevant sites and sites' relationships in proteins: Statistical Coupling Analysis (SCA)[13] and Direct Coupling Analysis (DCA).[7]

Let us start with the comparison with SCA, referring the interested reader to ref. 13 or to the ESI† for a detailed description. In brief, SCA is based on comparing the pairwise correlation matrix with the one built from a randomly reshuffled MSA with the same single amino acid frequencies at each position. Focusing on those principal components that correspond to eigenvalues that stand out of the distribution of the eigenvalues of the randomised matrix, one can define sectors corresponding to (putative) functional regions (see the ESI†). By applying SCA to RR we identified two relevant principal components. The projections of these components along the sequence allow us to represent each amino acid position in a two dimensional plane. The distance of each point from the origin can be taken then as a measure of relevance of the corresponding site, according to SCA.

Fig. 4(a) shows the overlap between the lists of the n most relevant sites according to CVS and SCA. For $n \leq 30$ we find that the overlap is smaller than what one would expect from random lists, this meaning that CVS and SCA are sensitive to different pieces of statistical information. This is not surprising: SCA is tailored to identify the most relevant drivers of variability in a dataset, and highly conserved sites which, as we saw for our *in silico* sequences, are effectively selected by CVS, do not appear on top of the list of relevant sites of SCA. This fact emerges more evidently by the analysis of the single site conservation as a function of the rank of the sites, according to the two methods (see Fig. 4(b)). The top sites selected by CVS have indeed much
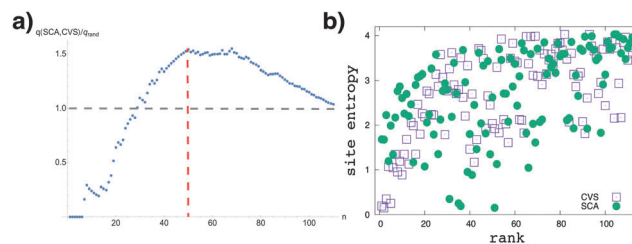

Fig. 4 (a) Overlap between the top n sites obtained by SCA and CVS (q(SCA,CVS)) normalised to the random value ($q_{\mathrm{rand}} = (n/L)^2 + (1 - n/L)^2$) as a function of the top n sites. For small n, this ratio is equal to zero. It increases afterwards, going beyond the random value for $n \simeq 30$ and it tends to q_{rand} for $n \to L$ (dashed grey line). In our comparisons, we took the value of n less than half the sequence ($n \simeq 50$) for which the overlap is maximal (red line). However, one can notice another maximum to be at $n = 68$. (b) Site entropy as a function of the rank according to CVS (squares) and SCA (circles).

** These reduced MSAs are obtained by picking sequentially at random the sequences of the original MSA and by including them in the new MSA if their minimal distance to the sequences already included is more than d. We checked that the outcomes are robust, in the sense that they do not depend on the specific realisation of this reweighing process.
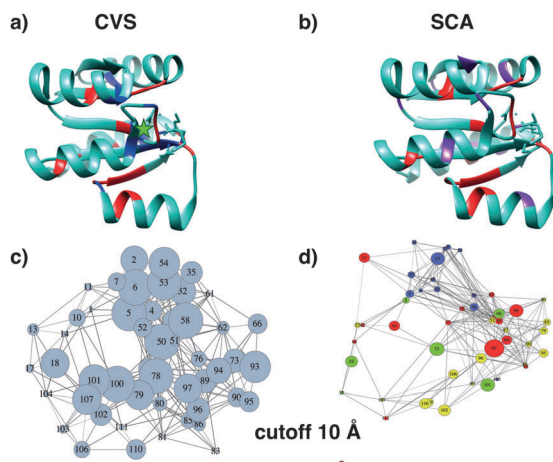
Fig. 5 (a and b) Top 30 sites (gapped position excluded) obtained by CVS (a) and SCA (b) visualised on the 3D structure of the response regulator receiver (PDB 1NXW). Common sites are highlighted in red, whereas sites singled out only by CVS (SCA) are marked in blue (purple). The active and phosphorylation site (52 on 1NXW) is marked in green in (a) and selected by CVS among the top 30 sites. (c and d) Network representation of the CVS top 50 sites (c) and of the SCA top 50 sites (d). The size of the nodes is rescaled respectively with the CVS counts C_i and the notion of relevance defined for SCA (*i.e.*, distance from the origin in the first two principal components space). The links between sites are given by the actual distances between the residues (cutoff 10 Å) on the 3D structure (PDB 1NXW), rescaled as well according to their own value. In (d) different colours represent the four sectors (see ESI†).

lower entropy than the corresponding ones for SCA. SCA and CVS are thus qualitatively very different methods. We can further visualise the difference between the two methods by depicting the 30 top sites identified by CVS and SCA on the protein tertiary structure. We analysed one structure (PDB ID 1NXW[21]) whose results are shown in Fig. 5(a) and (b). Some common domains on the α-helices and β-sheets are clearly identified by both methods (red bands). On this structure, a particular site, the Asn-52, has been identified as an active and phosphorylation site.[22,23] Most of the sites identified by only CVS (blue bands) lie indeed around Asn-52 (green star in Fig. 5(a)), whereas those identified only by SCA (purple bands) are much more scattered. In particular, the central β-sheet is singled out by CVS but not by SCA.

When one compares the lists for a larger number n of top sites, the overlap sharply increases. For $n = 50$ sites, the overlap between the two lists becomes maximal ($\sim$78%) as compared to what one would obtain if sites were randomly chosen ($\sim$45%).†† Still, the sites identified by CVS happen to be in closer spatial proximity on the three dimensional structure, with respect to those in the SCA list. In order to make this statement more precise, we pictured the two outcomes in a network structure. Each site among the top 50 identified by the two methods defines a node in such a network. The links between two sites are established by their proximity on the

†† Since sequence alignment procedures introduce gaps, we remove those sites that result in gaps in the consensus sequence. The number of selected gaps was very low though, less than 4% in both cases.

3D structure (PDB ID 1NXW) with a cutoff of 10 Å. The networks for CVS and SCA are represented in Fig. 5(c) and (d). Visual inspection reveals that CVS sites are more densely connected both in terms of the number of links and of the size of the largest connected component, and less fragmented, meaning that CVS sites interact more than those selected by SCA. We also notice that, for SCA, sectors do not seem to be related to a spatial pattern in the network (Fig. 5(b)). Summarising, we can conclude that CVS identifies a core of close sites in the protein, localised around the active site Asn-52.

We also applied SCA to the MSA of the voltage sensor domain. The results are comparable to those previously obtained for the RR. The overlap between the top n sites of SCA and CVS is smaller than the random threshold for $n < 35$, and it increases thereafter, showing a maximum for $n = 60$ sites, for which the overlap is about 80%.

A different way to visualise the difference between the two methods is to complement our analysis with Direct Coupling Analysis (DCA). DCA is a method aimed at identifying a network of interactions between positions along a protein domain, that are inferred from the traces left by the evolutionary process on the pairwise correlation matrix. In recent years many efforts have been spent in refining this observation into a quantitative bioinformatic tool.[6–12] Given the MSA of a certain protein family, DCA usually produces an F-score $F_{i,j}$ for each pair of positions $i, j = 1,\ldots,L$ with $F_{i,i} = 0$. In particular, if two positions are relevant for preserving the tertiary structure, by establishing a physical contact, one expects residues on these sites to co-evolve, resulting in a large value of $F_{i,j}$. DCA is indeed a powerful tool for predicting contacts in protein domains.

Here we use DCA to generate a network of interactions between positions, with the goal of deriving an independent assessment of the relevance of the sites selected by CVS and SCA, respectively. We stick to a standard implementation method of DCA – the so called naïve mean field direct coupling analysis (DCA) based on the so-called Plefka expansion[24] of statistical physics that has been shown to capture most of the direct contacts in 3D protein structures[7]. We refer to the ESI† for a concise discussion of the steps leading to the calculation of $F_{i,j}$ as applied to our dataset.

The m contacts with the largest value of $F_{i,j}$ define a network among the n top sites of CVS and SCA. A fraction of the m selected links will connect two of the n top sites of a given list, so the density of links and the fragmentation of the resulting networks provides insights into the nature of the co-evolutionary process taking place on the selected sites. Fig. 6(a) and (b) display the links of the $m = 60$ top contacts that connects two of the $n = 50$ most relevant sites according to CVS (a) and SCA (b) sites, respectively. Interestingly, the putative contacts singled out for CVS turn out to correspond to closer sites with respect to those identified for SCA. In general, the sites selected by CVS appear to be more densely connected than those identified by SCA. Indeed, by building up a network using as nodes the top n sites selected by respectively CVS and SCA and as links the top m DCA contacts, one can measure the number of actual contacts, K_{CVS} and K_{SCA}, and the size of the largest connected component
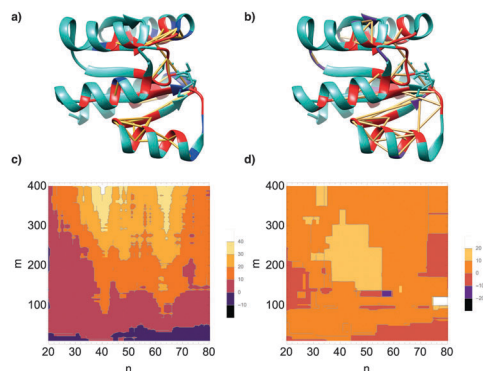
Fig. 6 (a and b) Top 60 contacts amplified by DCA reduced on the sublists of the top 50 sites highlighted on the 3D structure (PDB 1NXW) for CVS (a) and SCA (b). (c) Density plot of the difference $K_{CVS} - K_{SCA}$ between the number of links in the subnetwork of the n top sites of CVS and SCA, as a function of n and the number of DCA contacts m. Here K_L is the number of contacts, among the m top contacts identified by DCA, connecting sites belonging to the list L = CVS, SCA. (d) Difference between the size of the largest connected component of the sub-networks of the n top sites of CVS and SCA identified by the top m DCA contacts.

Table 1 Comparison with databases for CVS and SCA results on RR. Functional sites have been extracted from ref. 22,23 for the B4DA37_9BACT sequence and[18] referring to the structure PDB ID 2CHF. In order to make a consistent comparison sequences have been matched with particular attention to the gaps

Source	Our alignment	Function	Rank SCA	Rank CVS
22 and 23	5	Active	20	3
22 and 23	6	Active	31	2
18	47	Binding site for phosphate	83	87
18	50	Phosphorylation	36	1
22 and 23	52	Active, phosphorylation	25	31
22 and 23	55	Intermolecular recognition site	78	86
22 and 23	56	Intermolecular recognition site	18	85
22 and 23	58	Intermolecular recognition site	35	5
22 and 23	59	Intermolecular recognition site	62	84
22 and 23	60	Active, intermolecular recognition site	59	83
18	75	Hydrogen bond	98	61
22 and 23	80	Active	44	37
18	97	Phosphorylation	1	14
22 and 23	99	Active	48	62
22 and 23	102	Active and dimerization interface	8	25
22 and 23	103	Active and dimerization interface	75	43
22 and 23	104	Active and dimerization interface	64	46

in the two cases. This allows us to notice that that the top n sites identified by CVS share a larger number of the m top DCA contacts and own a larger connected component, with respect to those identified by SCA (see Fig. 6(c and d) and figure caption), in a wide range of values of n and m. Furthermore, the number of links between the n top sites and the remaining $L - n$ sites, is smaller for CVS than for SCA (not shown). This is an additional indication of the sharp separation between relevant and irrelevant sites afforded by CVS.

4.3 Biological relevance

In order to further assess the relevance of sites selected by CVS, we analysed their biological relevance comparing them to functional sites stored in databases, the solvent accessible surface rate[25,26] and their importance in the evolutionary dynamics.

4.3.1 The response regulator receiver. We extracted functional sites from the UniProtKB and NCBI-RefSeq databases[22,23] and from ref. 18 and checked their rank in the hierarchies defined by CVS and SCA. As shown in Table 1, most of the functional sites are ranked by both methods among the top 50, showing the ability of both methods of capturing functional features. Remarkably CVS identifies as top 3 sites two active residues (5, 6) and the previously mentioned active and phopshorylation site (50). An important functional domain of this protein, the intermolecular recognition domain, contains one site ranked as 5th by CVS (58) and its surrounding sites are successively ranked, although captured later on. The same behaviour can be noticed within the dimerisation interface (residues 102, 103, and 104).

Another important biological aspect to investigate concerns the position of the relevant sites on the 3D structure. To this aim, we analysed how the total CVS counts correlate with the solvent accessible surface rate (SAS).[25,26] The SAS provides a measure of the accessibility of a given site in the 3D protein structure, to small molecules such as water. Sites with large SAS values are typically on the outer surface of the protein whereas those with a low SAS value are buried in the interior. We compared the distributions of the SAS of the n most relevant sites identified by CVS and SCA with the overall distributions as well as the z-score of the top n sites identified by the two methods defined as $z\text{-score} = \dfrac{\langle SAS_x \rangle - \langle SAS \rangle}{\delta SAS}$, where $\langle SAS_x \rangle_n$ is the average SAS of the top n sites identified by x = CVS, SCA, $\langle SAS \rangle$ is the average SAS over all the sites and δSAS is its standard deviation. Fig. 7(b) shows that the sites selected by CVS have a consistent bias towards lower values of the SAS, indicating that CVS preferably selects internal and conserved sites, which are putatively important for the maintenance of structural properties of the protein. The sites selected by SCA are also preferentially interior, but the bias is considerably weaker.

Finally, as discussed in the Introduction, these sequences all come from a common evolutionary history, being selected as the optimal solution to carry out the same function across
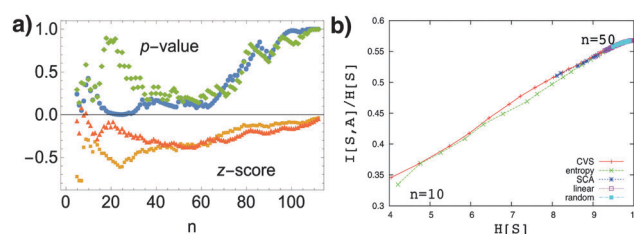

Fig. 7 (a) p-Value obtained from the Kolmogorov−Smirnov test between the Solvent accessible surface distributions of the top n relevant sites obtained by CVS (blue circles) and SCA (green diamonds) and the overall one along with the corresponding z-score as defined in the text for CVS (orange squares) and SCA (red triangles). Solvent accessible surface rates have been computed using Naccess[25,26] with PDB 1NXW as the input. (b) Mutual information about the sequences and the annotations as defined in the text for different ranking methods (the legend is shown as the inset).

different organisms. We here asked then whether the sites identified by CVS can be the carriers of the relevant evolutionary dynamics. To assess this point, we computed the mutual information between the subsequences identified by different methods and the annotation that identifies the organism of origin for each sequence, given by the mnemonic organism identification code of UniProtKB. We compare the lists of the top rank n sites obtained according to CVS, SCA and just simple conservation (*i.e.* the n positions with the smallest site entropy) among themselves and with a list of randomly chosen sites. Different lists afford different levels of variation (*i.e.* different subsequence entropies $H[\underline{s}]$), so a meaningful comparison is not at constant n but at constant $H[\underline{s}]$. The fraction of this variation, that accounts for the variation between organisms, was found to be largest for CVS compared to both SCA and conservation, for $n \in [10, 50]$ (see Fig. 7(b)).

This analysis lets us conclude that CVS is capable of extracting functional sites with low SAS along with more conserved sites, usually internal in the protein 3D structure, as well as accounting for sequence variation across different organisms.

4.3.2   The voltage-sensor domain of the ion channels. To further probe the validity of our results, we studied the voltage-sensor domain of the ion channel sequences. We investigated the biological relevance of CVS relevant sites, *i.e.*, functionality and solvent accessible surface rates. By running our algorithm $R = 100$ times for each subsequence length ($n = 5, 10, 15, ...,60$), one is able to build a full count statistics. On the full dataset of $N = 6652$ sequences, we found a dispersion smaller than 10%. In order to test the stability, we run it also on a subset of

$N = 666$ sequences for $n = 40$ we found a larger overlap of 76% with the counts obtained for the full dataset.

As shown in Fig. 8(d), by ranking the counts, CVS also in this case clearly distinguishes relevant from irrelevant sites affording a sharp division between these two sets. Fig. 8(e) shows that, besides highly conserved sites, CVS distinguishes between sites whose variability is evolutionarily related to those that can be regarded as noise. By visualising the top ranked sites on the 3D protein structure as well as comparing these results with the available functional knowledge, among these top sites, a first group of 15 positions with high counts C_i can be identified. This contains 9 sites identified in ref. 17 and 27. Three more functionally relevant sites have counts larger than 500 belong to a larger group of the 38 most relevant sites. These sites are represented in the 3D structure in Fig. 8(b and c). These include N-62, N-72, R-76 and E-93 of the voltage-dependent K^+ channel KvAP which are important for channel function.[23,27] The same sites have also been identified in ref. 17 that refer to the NavAb sequence (E-49, E-59, R-63, D-80 respectively). Ref. 17 also highlights the role of I-22, F-56 and F-71 in NavAb and it discusses the application of direct coupling analysis of the VSD MSA, identifying several evolutionary conserved contacts along the chain. In particular, E-49 is found to be in contact both with N-25 and with E-96, which are far apart in the NavAb structure. Ref. 17 argues that these two contacts are important to confer stability both to the activated and to the resting state of the protein domain. All these sites are found to be relevant in the CVS analysis, as well as R-63 and S-77, which are also found to be in contact with the NavAb structure.
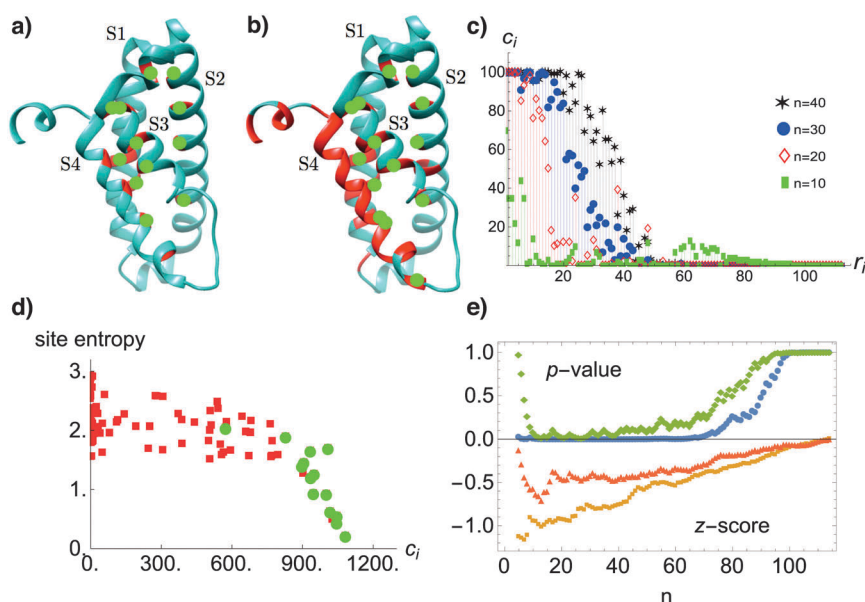

Fig. 8 (a and b) Top CVS 15 (a) and 40 (b) sites represented on the VSD 3D structure (PDB ID 3RVZ) obtained by running CVS on the dataset 100 times for each subsequence length $n = 10, 20, 30$ and 40 (respectively, green squares, empty red diamonds, blue circles and black stars) and then ranking sites according to the total count, C_i. Green circles spot the functional sites already identified in ref. 23 and 27 as discussed in the main text. (c) Ranked relevance count $c_i(n)$ for positions in VSD for different subsequence lengths, *i.e.* $n = 10, 20, 30$ and 40. (d) Site entropy as a function of total count. Green circles represent positions identified in the literature (see text) represented on the 3D structure in (b) and (c). (e) p-Value obtained from the Kolmogorov–Smirnov test between the solvent accessible surface distributions of the top n relevant sites obtained by CVS (blue circles) and SCA (red squares) and the overall one. Solvent accessible surface rates have been computed using Naccess[25,26] with PDB 3RVZ as the input.

Ref. 17 also reports a false positive contact (between W-76 and T-15). We find that while W-76 is relevant, T-15 is not ($C_{T\text{-}15} = 145$). Finally, we find an enrichment in relevant sites in the region corresponding to S4, which is a highly dynamical region of the VSD, and in the S2–S3 turn (Y-63 to P-95 in KvAP) that has been suggested to be structurally important.[27]

We then proceeded again by comparing the solvent accessible surface rate distributions on the top n sites identified by CVS and SCA respectively (see Fig. 8(e)): in this case, the outcome of the two methods is definitely different up to the top 15 sites, becoming more similar beyond this threshold, in agreement with what obtained before for the RR. Again, we find a bias towards internal sites, which is stronger for CVS than for SCA. The example of the VSD sample shows that, in spite of the moderate size of the MSA, CVS is able to extract stable results and to identify functionally relevant sites.

5 Conclusions

In summary, we have proposed a new method for the identification of a core of functionally and structurally relevant sites in protein domains. Given a multiple sequence alignment (MSA) for a protein family, the method is based on finding those subsets of n positions for which the entropy $H[K]$ of the frequency with which different subsequences occur is maximal, corresponding to broad frequency distributions for these subsequences. Its implementation is straightforward and it does not require any further data-processing step. By starting with subsequences of different lengths, the method assigns to each position a count that is used to assess the relevance of a certain site. As the subset length includes enough positions, CVS affords a sharp separation between relevant and irrelevant sites for all the datasets we analysed, the relevant sites are often but not always highly conserved positions. Besides, site relevance turns out to increase with the subset size n, *i.e.*, typically, if a site i is selected as relevant in the subset of the n most relevant positions, it will be very likely selected in the subset of $n' > n$ sites as well.

The application of our method to *in silico* sequences provides a first check of the ability of CVS for discriminating relevant information from noise and capturing dependencies going beyond pairwise correlation in big datasets.

We have discussed the application of the method to two protein domain families, the response regulator receivers (RR, PFAM ID PF00072) and the voltage sensor domain of the ion channels (VSD, PFAM ID PF00520). We first studied the response regulator receivers (RR) and inspected the robustness of the method against reweighing and its ability for going beyond pure single site conservation and pairwise correlations. We then compared our method with statistical coupling analysis (SCA) by analysing the RR and the VSD. After assigning a measure of relevance to both methods, we studied the overlap between the two solutions finding out that although the top most relevant sites are normally quite different, this overlap increases. For the top 50 sites in the RR, the overlap between the two solutions is 78%.

An analogous result was found for the VSD. Yet, the small differences between CVS and SCA results can be furthermore highlighted by using Direct Coupling Analysis (DCA). CVS identifies a core of densely connected residues and all significant contacts predicted by DCA on the restricted lists turn out to be close on the 3D structure for the RR. CVS is furthermore capable of identifying biologically relevant positions: the sites extracted from ref. 22,23,27 or identified by inference methods as direct coupling analysis turn out to be tagged as relevant by our method for both RR and VSD. This further corroborates the conclusion that CVS indeed singles out subsets of relevant sites in protein domains.

We stress, in particular, the fact that CVS is able to recover insights from methods based on single site conservation and pairwise correlation. Yet, the most exciting aspect of CVS lies precisely in its ability to probe the co-evolutionary process beyond single site conservation and pairwise correlation. This calls, on the one hand, for the development of inference methods going beyond pairwise interactions, and on the other hand to applications of CVS to instances that may lead to a more critical assessment of its potential for reverse engineering evolutionary processes.

References

1 J. M. Berg, J. L. Tymoczko and L. Stryer, in *Biochemistry*, ed. W. H. Freeman, 5th edn, 2002.
2 J. Overington, M. S. Johnson, A. Sali and T. L. Blundell, *Proc. R. Soc. B*, 1990, **241**, 132–145.
3 A. A. Fodor and R. W. Aldrich, *J. Biol. Chem.*, 2004, **279**, 19046–19050.
4 M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner and R. Ranganathan, *Nature*, 2005, **437**, 512–518.
5 B. Lunt, H. Szurmant, A. Procaccini, J. A. Hoch, T. Hwa and M. Weigt, *Methods Enzymol.*, 2010, **471**, 17–41.
6 M. Weigt, R. A. White, H. Szurmant, J. A. Hoch and T. Hwa, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 67–72.

7 F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa and M. Weigt, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, E1293–E1301.

8 S. Cocco, R. Monasson and M. Weigt, *PLoS Comput. Biol.*, 2013, **9**, e1003176.

9 M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt and E. Aurell, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2013, **87**, 012707.

10 C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt and A. Pagnani, *PLoS One*, 2014, **9**, e92721.

11 C. Feinauer, M. J. Skwark, A. Pagnani and E. Aurell, *PLoS Comput. Biol.*, 2014, **10**, e1003847.

12 M. Ekeberg, T. Hartonen and E. Aurell, *J. Comput. Phys.*, 2014, **276**, 341–356.

13 N. Halabi, O. Rivoire, S. Leibler and R. Ranganathan, *Cell*, 2009, **138**, 774–786.

14 M. Marsili, I. Mastromatteo and Y. Roudi, *J. Stat. Mech.: Theory Exp.*, 2013, **2013**, P09003.

15 A. Haimovici and M. Marsili, *J. Stat. Mech.: Theory Exp.*, 2015, P10013.

16 W. Ewens, *Theor. Popul. Biol.*, 1972, **3**, 87–112.

17 E. Palovcak, L. Delemotte, M. L. Klein and V. Carnevale, *J. Gen. Physiol.*, 2014, **143**, 145–156.

18 A. H. West and A. M. Stock, *Trends Biochem. Sci.*, 2001, **26**, 369–376.

19 W. A. Catterall, *Neuron*, 2010, **67**, 915–928.

20 K. A. Reynolds, W. P. Russ, M. Socolich and R. Ranganathan, *Methods Enzymol.*, 2013, **523**, 213–235.

21 R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate and M. Punta, *Nucleic Acids Res.*, 2014, **42**, D222–D230.

22 T. U. Consortium, *Nucleic Acids Res.*, 2014, **42**, D191–8.

23 The NCBI handbook [Internet], *National Center for Biotechnology Information*, National Library of Medicine (US), Bethesda (MD), 2002, Ch. 18, The Reference Sequence (RefSeq) Project. Available from http://www.ncbi.nlm.nih.gov/books/NBK21091/.

24 T. Plefka, *J. Phys. A: Math. Gen.*, 1982, **15**, 1971.

25 B. Lee and F. Richards, *J. Mol. Biol.*, 1971, **55**, 379–400.

26 S. J. Hubbard and J. M. Thornton, *NACCESS, Computer Program*, Department of Biochemistry and Molecular Biology, University College London, 1993, http://bioinf.manchester.ac.uk/naccess/.

27 S. Lee, A. Lee, J. Chen and R. MacKinnon, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15441–15446.