PLoS one

Integrated Analysis of Residue Coevolution and Protein Structure in ABC Transporters

Attila Gulyás-Kovács*

Laboratory of Cardiac/Membrane Physiology, Rockefeller University, New York, New York, United States of America

Abstract

Intraprotein side chain contacts can couple the evolutionary process of amino acid substitution at one position to that at another. This coupling, known as residue coevolution, may vary in strength. Conserved contacts thus not only define 3-dimensional protein structure, but also indicate which residue-residue interactions are crucial to a protein's function. Therefore, prediction of strongly coevolving residue-pairs helps clarify molecular mechanisms underlying function. Previously, various coevolution detectors have been employed separately to predict these pairs purely from multiple sequence alignments, while disregarding available structural information. This study introduces an integrative framework that improves the accuracy of such predictions, relative to previous approaches, by combining multiple coevolution detectors and incorporating structural contact information. This framework is applied to the ABC-B and ABC-C transporter families, which include the drug exporter P-glycoprotein involved in multidrug resistance of cancer cells, as well as the CFTR chloride channel linked to cystic fibrosis disease. The predicted coevolving pairs are further analyzed based on conformational changes inferred from outward- and inward-facing transporter structures. The analysis suggests that some pairs coevolved to directly regulate conformational changes of the alternating-access transport mechanism, while others to stabilize rigid-body-like components of the protein structure. Moreover, some identified pairs correspond to residues previously implicated in cystic fibrosis.

Citation: Gulyás-Kovács A (2012) Integrated Analysis of Residue Coevolution and Protein Structure in ABC Transporters. PLoS ONE 7(5): e36546. doi:10.1371/journal.pone.0036546

Editor: I. King Jordan, Georgia Institute of Technology, United States of America

Received February 7, 2012; **Accepted** April 6, 2012; **Published** May 8, 2012

Funding: This study was supported from National Institutes of Health (NIH) research grant NIH DK51767 (to D. C. Gadsby). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: agulyaskov@rockefeller.edu

Introduction

The increasing number of solved protein structures raises the question how structural data can help clarify the biochemical mechanisms underlying protein function. Although extremely informative, even the complete map of residue contacts is in general insufficient to reveal biochemical mechanisms. Experiments mutating specific amino acid positions are essential complements to structure but the typically low throughput of these experiments calls for highly specific, rational design. Sometimes structural models themselves highlight experimental candidate positions but more often additional information is needed. This is especially so when specific functional interactions, represented by pairs of positions, are to be tested [1,2] since the number of candidate pairs scales, in principle, as the square of the number of candidate positions.

The superfamily of ATP-binding cassette (ABC) transporters is an epitome of proteins with recently determined structures but poorly understood biochemical mechanisms [3,4]. Their members actively transport substrate molecules across membranes with the exception of the (passive) ion channel CFTR (a member of the ABC-C family), whose defect causes cystic fibrosis disease. Typical members of the ABC-B and ABC-C families are active exporters, like the MDR and MRP proteins (notably Pgp/MDR1), which recognize anticancer drugs as their natural substrates and thereby confer multidrug resistance on tumor cells.

All ABC-B and ABC-C transporters are built of two transmembrane domains (TMDs), which interact directly with the translocating substrate, and two nucleotide binding domains (NBDs), which convert chemical to mechanical energy by binding and hydrolyzing ATP (Figure 1A). The popular alternating-access transport model asserts that this mechanical energy drives a conformational cycle coupled to unidirectional transport, and during each cycle the TMDs alternate between inward and outward-facing conformation [5]. This model, although supported by relatively high-resolution structures [3,4], describes transport mechanism at a resolution that is too low for the clarification of many crucial details related to multidrug resistance or cystic fibrosis. For a refined model, mechanistically crucial residue-residue interactions need to be somehow predicted and experimentally tested: particularly between the transmembrane helices (TM1,TM12), which are relatively understudied, and whose extensions form intracellular loops (ICL1,ICL4), which couple the TMDs to the NBDs (Figure 1A).

The abundance of sequenced ABC-B and ABC-C proteins makes these families ideal for comparative sequence analysis. Such analysis can infer those structural and functional constraints on sequence evolution that are not necessarily evident from sole structural analysis. For example, side chain contacts can couple the process of amino acid substitution at one position to that at the contacting position and thereby induce residue coevolution, but the strength of coupling and its persistence in time may vary [6,7].

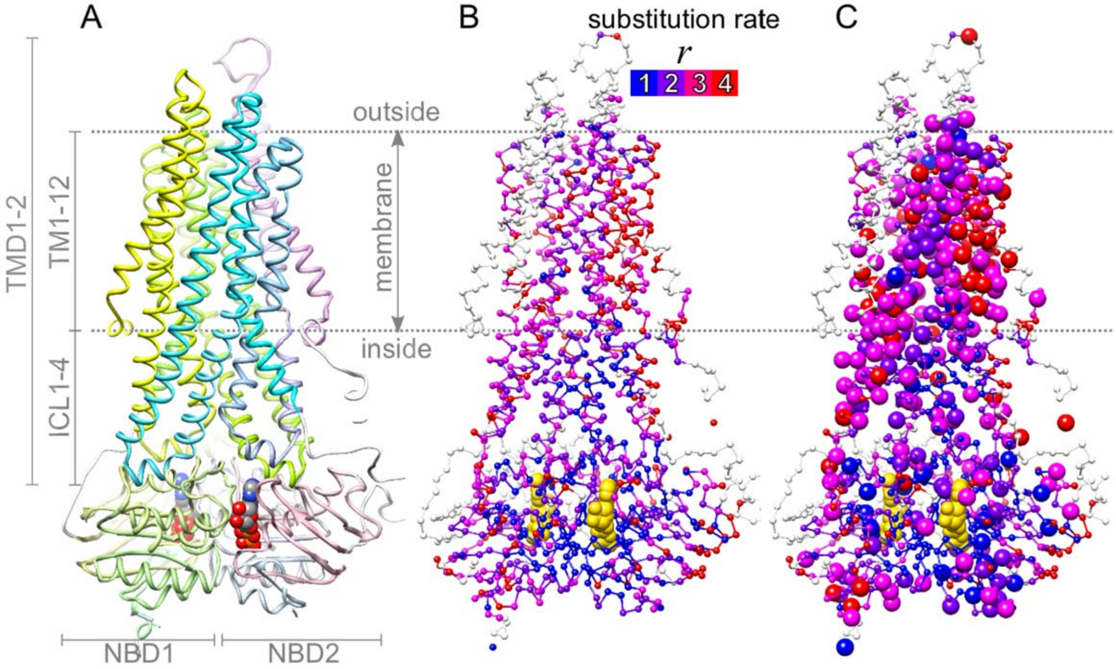

Figure 1. Structure of ABC-C proteins and the rate of amino acid substitution. (**A-C**) Homology model of the ABC-C protein CFTR [45]. (**A**) Main structural components. *NBD*: nucleotide binding domain; *TMD*: transmembrane domain; *TM*: transmembrane helix; *ICL*: intracellular loop. ATP-molecule atoms are shown as spheres. (**B**) Each amino acid position i is marked by a small sphere at the C_α atom and is colored according to r^i, the estimated discretized substitution rate (eq. 21). $r^i = 1$ (blue) indicates that i is conserved. (**C**) The large spheres represent the set of positions predicted in this study to coevolve with some other position(s) in the same set. Structural figures were made using UCSF Chimera [70].
doi:10.1371/journal.pone.0036546.g001

Therefore, statistical techniques predicting coevolving pairs, henceforth referred to as coevolution detectors, have been utilized for different purposes. When the representative structure of some protein family is unknown, then coevolution detectors can be used to predict contacts and thereby aid structure determination [8–16]. But when such structure is known, detectors are still useful for the prediction of the subset of contact pairs that exhibit strong and permanent coevolution [11,17–25]. The latter set of pairs can be interpreted as a representation of conserved and general mechanisms that characterize the whole protein family. Therefore, these pairs are highly relevant for the elucidation of these mechanisms as either self-standing results or pointers for the rational design of "double mutants" [1,2,26–28] for functional experiments.

All coevolution detectors predict coevolving pairs from multiple sequence alignments but they differ from each other in crucial assumptions on the substitution process, which can profoundly affect prediction accuracy. Yet the relative performance of individual detectors in accuracy tests remains unclear even after side by side comparison [29,30], suggesting that accuracy strongly depends on the specific protein family and certain properties of the corresponding alignment. Therefore, a key question is: given a collection of detectors and a protein family with representative sequences and structure(s), how can coevolving pairs be detected the most accurately?

The present study addresses that question with a new, integrative framework (Figure 2), which improves accuracy by directly incorporating structural information and by combining multiple detectors. Moreover, it features procedures that deal with the well-known vulnerability of detectors to the statistical non-independence of homologous sequences [31–33] and to the heterogeneity of positions with respect to substitution rate [34,35]. This framework is employed to ABC-B and ABC-C transporters to predict those contact pairs that represent evolutionarily conserved interactions (i.e. coevolving pairs). The predicted pairs are presented with a particular attention to the possible mechanistic coupling between TM helices in both the inward and outward conformation of the TMDs.

Methods

Central Assumptions of the New Framework

Considering pairs of amino acid positions in a protein family, assume that, for each pair, the two positions either strongly and permanently coevolve with each other or evolve completely independently. Let E denote the set of *coevolving* pairs. Let S represent the set of (structural) *contact* pairs, specifically side chains contacts. Following pioneering studies [13,14,16] an intimate relationship has been conjectured between coevolution and side chain contact. The relationship can be stated in terms of the probabilities $\Pr(E)$ and $\Pr(E|S)$ that, for some protein family, a random draw from all pairs or from contact pairs, respectively, gives a coevolving pair:

$$\Pr(E) < \Pr(E|S). \quad (1)$$

This says that the contact pairs tend to be the coevolving pairs. Let P be the set of coevolving pairs *predicted* by some coevolution detector from sequence data D. If the detector is useful then conditioning on P has similar effect to conditioning on S:

$$\Pr(E) < \Pr(E|P). \quad (2)$$

Supporting the preceding two assertions it has been shown repeatedly [11–14,16,20,22,23,29–32,35–42] that most detectors

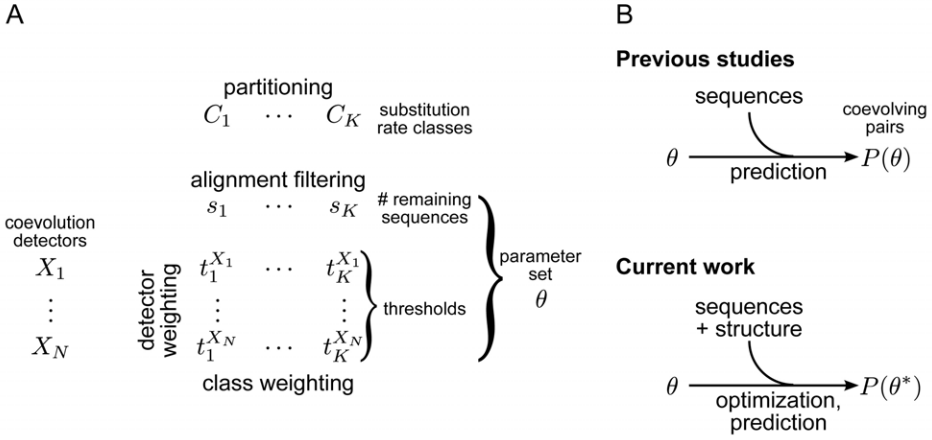

Figure 2. Integrative framework for the prediction of coevolving position pairs. (**A**) Parameters of the framework, and weighting and filtering procedures controlling them. *Partitioning* the set of all position pairs into substitution rate classes C_k (eq. 10, 20–22), and *weighting* each *class* (eq. 11–13), addresses the sensitivity of coevolution detectors to substitution rate. *Detector weighting*: previous studies employed coevolution detectors X_n either separately or in a combination $X_1 \wedge \ldots \wedge X_N$ in which all X_n were equally weighted. However, equal weighting of $X_1 \wedge \ldots \wedge X_N$ is not generally the optimal combination as demonstrated below in Figure 3B. The new framework allows unequal weighting of detectors (eq. 15). *Alignment filtering* (eq. 9, 14) removes redundant sequences from the input data (the sequence alignment) to minimize the adverse influence of phylogenetic redundancies on detectors. (**B**) Previous studies predicted coevolving position pairs in a protein family from only the corresponding sequence alignment, while ignoring useful information in solved structures. The current work makes use of structural information to adjust the parameters of detector weighting, class weighting and alignment filtering (parameter set θ) for optimal performance, as gauged by prediction of known structural contacts (eq. 5, 19).

doi:10.1371/journal.pone.0036546.g002

can predict contact pairs better than random choice, and so

$$\Pr(S) < \Pr(S|P). \tag{3}$$

Instead of predicting contact pairs to aid de novo prediction of structure, several studies [11,18–25] aimed to detect coevolving pairs given the set of contact pairs assuming that

$$\Pr(E|S) < \Pr(E|S \cap P). \tag{4}$$

The new framework was designed towards that aim and takes all above assumptions and findings as a starting point. As Figure 2 shows, P depends on a set of parameters θ, which specifies the identity of the detector (when a single detector is used) or the relative weights of detectors (when multiple detectors are combined). θ also determines how data are analyzed by a given (set of) detector(s): how classes of pairs are weighted and how the input alignment is filtered (Figure 2A). Therefore, if the protein structure is known, then θ can be adjusted for optimal prediction of contact pairs. The individual parameters and the optimization problem will be precisely stated later; at this point another possible formulation is given to be consistent with eq. 3:

$$\theta^* = \arg\max_{\theta} \Pr(S|P(\theta)). \tag{5}$$

A crucial assumption of this study is that the optimization in eq. 5 improves the detection of coevolving pairs within the set of contact pairs:

$$\Pr(E|S \cap P(\theta)) \leq \Pr(E|S \cap P(\theta^*)). \tag{6}$$

Thus the central goal of this work is to find θ^*, which uniquely determines $P(\theta^*)$ (Figure 2B) and ultimately $S \cap P(\theta^*)$. A key feature of the new framework is that the known structure plays a dual role in the current analysis. First, the structure is required for the optimization of the parameters (Eq. 5, Figure 2B bottom). Second, the structure (or some alternative conformation of that structure) is used to restrict the predicted pairs to the set of contact pairs by taking the intersection $S \cap P$ (Eq. 6).

Parameters and Procedures of the New Framework

As mentioned above, P is a function of the parameter set θ. Now the question is: exactly what is θ, and how does it determine P together with the data?

In general, a coevolution detector X acts as a binary classifier that divides the set Ω of all pairs into P and the complementary set of pairs (the "negatives"). Given the input alignment data D, the condition for classification of each pair p into P is that the test statistic T^X of the detector evaluated at p exceeds an adjustable threshold t:

$$P(t,D) = \{p \in \Omega | T^X(p,D) \geq t\}. \tag{7}$$

It is practical to constrain the number of predicted pairs $|P|$ at some chosen fraction γ of all pairs by treating t as a monotonically increasing function of γ. Then, for a given X and D,

$$P(\gamma) = \{p \in \Omega | T(p) \geq t(\gamma)\}, \quad |P| = \gamma|\Omega|. \tag{8}$$

Consequently, γ controls the true and false positive rate of the detector, which are defined subsequently in eq. 16–17.

The procedure of *filtering* of an alignment of homologous sequences, in particular *phylogenetic* type of filtering, aims to remove redundancies that emerge from the statistical non-independence within any collection of homologous sequences. These redundancies pose challenges to all coevolution detectors, especially to those assuming that homologous sequences are statistically independent from each other.

Any type of filter, applied to alignment D, permutes sequences in a given order that depends on the filter type F. Then the filter removes a certain number of sequences in that order. Therefore, the filtered D is determined both by F and by the number s of sequences that remain in the alignment. It follows that, for a given X,

$$P(t,D,F,s)=\{p\in\Omega|T(p,D(F,s))\geq t\}. \tag{9}$$

Filtering will be discussed in more detail in Methods: Alignment Filtering.

For all detectors, $T(p)$ is known [34,35,38] to depend to some degree not only on the coevolution of position i and j (where $p=(i,j)$) but also on the overall rate of amino acid substitution at i and at j. The dependence on substitution rate deteriorates the performance of the detector but can, in theory, be addressed by conditioning t on the rates of the pair. Therefore, the new framework incorporates a novel strategy based on the procedure of *partitioning* Ω into K (substitution) *rate classes* C_k (Figure 2A):

$$\Omega=\bigcup_{k=1}^{K} C_k. \tag{10}$$

The precise definition of $\{C_k\}$ will be given later (eq. 20–22), but it may be worth emphasizing at this point that the members of each C_k are position *pairs* and not single positions. Now a key feature of the new framework is that t_k can be adjusted separately for each C_k and that P is defined as the union of the resulting P_ks:

$$P_k(t_k)=\{p\in C_k|T(p)\geq t_k\} \tag{11}$$

$$P=\bigcup_{k=1}^{K} P_k. \tag{12}$$

The vector $\mathbf{t}=(t_1,\ldots,t_K)$ thus determines every P_k and therefore every $|P_k|$. Like its scalar analog t, $\mathbf{t}$ is also a function of γ, which imposes the constraint

$$\sum_{k=1}^{K}|P_k(t_k)|=\gamma|\Omega|. \tag{13}$$

(This is the same as the constraint expressed by the second equality in eq. 8, since P_ks are disjoint sets and thus $|P|=\sum_k |P_k|$.) The constraint in eq. 13 still allows individual t_ks to vary, which changes the relative size (the weights) of P_ks. In this work the procedure of changing $\mathbf{t}$, while requiring eq. 13 to hold, is referred to as *class weighting* procedure.

Partitioning Ω also allows the filtering of D separately for each rate class so that there is a separate parameter s_k for each C_k,

$$P_k(t_k,s_k)=\{p\in C_k|T(p,D(s_k))\geq t_k\}, \tag{14}$$

and thus $P\equiv\bigcup_k P_k$ also depends on the vector $\mathbf{s}=(s_1,\ldots,s_k)$. Eq. 14 corresponds to the combination of *partitioning* + *class weighting* + *filtering* in case of a general $\mathbf{t}$ satisfying eq. 8, or to the combination of *partitioning* + *filtering* when all t_ks are set to the same value. Note that in this case "combination" refers to *procedures* and not *detectors*.

Up to this point a single detector X was assumed. Now let $\{X_n\}$ be a collection of N detectors, and let $X_1\wedge\ldots\wedge X_N$ denote their logical AND combination [43] and $\tau=(t^{X_1},\ldots,t^{X_N})$ the corresponding thresholds (Figure 3A). Then the set of pairs predicted by the combined detector $X_1\wedge\ldots\wedge X_N$ is defined as

$$P(\tau)=\{p\in\Omega|T^{X_1}(p)\geq t^{X_1},\ldots,T^{X_N}(p)\geq t^{X_N}\}. \tag{15}$$

It is clear that τ uniquely determines $|P|$ and that, for a given γ, the constraint $|P|=\gamma|\Omega|$ allows individual t^{X_n}s to vary. For some $1\leq m\leq N$, the impact of X_m on P, relative to that of any other detector X_n ($n\neq m$), increases with t^{X_m}. In other words, the weight of X_m increases in $X_1\wedge\ldots\wedge X_N$. Therefore, adjusting t^{X_n}s relative to each other is referred to as the procedure of *detector weighting* and is illustrated by Figure 3A.

Given a specific detector X_m, if $t^{X_n}\to-\infty$ for all other detectors X_n ($n\neq m$), then the weight of these detectors vanish. This special case is equivalent to using detector X_m alone and not in combination with other X_ns. Furthermore, in the general case it is straight-forward to combine *detector weighting* with *partitioning* + *class weighting* (Figure 2A). Then each scalar t^{X_n} is replaced by a vector $\mathbf{t}^{X_n}\equiv(t_1^{X_n},\ldots,t_K^{X_n})$ so that $\tau=(\mathbf{t}^{X_1},\ldots,\mathbf{t}^{X_N})$. This can be further extended with *filtering*.

In summary, given the parameter γ, data D, a filter type F, substitution rate classes $\{C_k\}$ and a set $\{X_n\}$ of detectors, the collection of parameters $\theta=(\tau,\mathbf{s})$ uniquely determines the set of predicted pairs $P(\gamma,\theta)$ in the new framework. Next, it will be discussed how the optimal θ^* is actually found, and eq. 5 will be replaced by a closely related formula. This will be followed by detailed information on $D,F,\{C_k\}$ and $\{X_n\}$.

Optimization Using Structural Information

Let $D,F,\{C_k\},\{X_n\}$ and $P(\gamma,\theta)$ have the same meaning as before. Let S denote the set of contact pairs and B the set of pairs $p=(i,j)$ for which i and j are separated by some substantial distance in 3D space, so that i and j are unlikely to directly interact with each other in any native conformation of the protein. S and B will be defined in the next subsection; for now assume that these sets are known. The true positive rate ρ^{TP} (sensitivity) and false positive rate ρ^{FP} (reverse specificity) are defined, respectively, as

$$\rho^{\mathrm{TP}}(\gamma,\theta)=\frac{|P(\gamma,\theta)\cap S|}{|S|}, \tag{16}$$

$$\rho^{\mathrm{FP}}(\gamma,\theta)=\frac{|P(\gamma,\theta)\cap B|}{|B|}. \tag{17}$$

As noted after eq. 8, ρ^{TP} and ρ^{FP} are functions of γ, and therefore eq. 16–17, together with eq. 8, shows that $\gamma\to0$ makes both ρ^{TP} and $\rho^{\mathrm{FP}}\to0$. Likewise, $\gamma\to1$ drives both ρ^{TP} and $\rho^{\mathrm{FP}}\to1$. In general, $\rho^{\mathrm{TP}}\neq\rho^{\mathrm{FP}}$ for a given detector and θ. When $\rho^{\mathrm{TP}}>\rho^{\mathrm{FP}}$, the detector is informative with respect to random selection. In contrast, for a theoretical *random detector* $\rho^{\mathrm{TP}}=\rho^{\mathrm{FP}}$ (Figure 3B-C, dashed line).

The receiver operator characteristic curve of a detector is a mapping that associates each γ with $(\rho^{\mathrm{FP}}(\gamma),\rho^{\mathrm{TP}}(\gamma))$ at a fixed θ (Figure 3B-C). The partial area $A(\alpha,\theta)$ under the ROC curve is the Riemann-Stieltjes integral of ρ^{TP} with respect to ρ^{FP} over the interval $[0,\alpha],(0\leq\alpha\leq1)$:

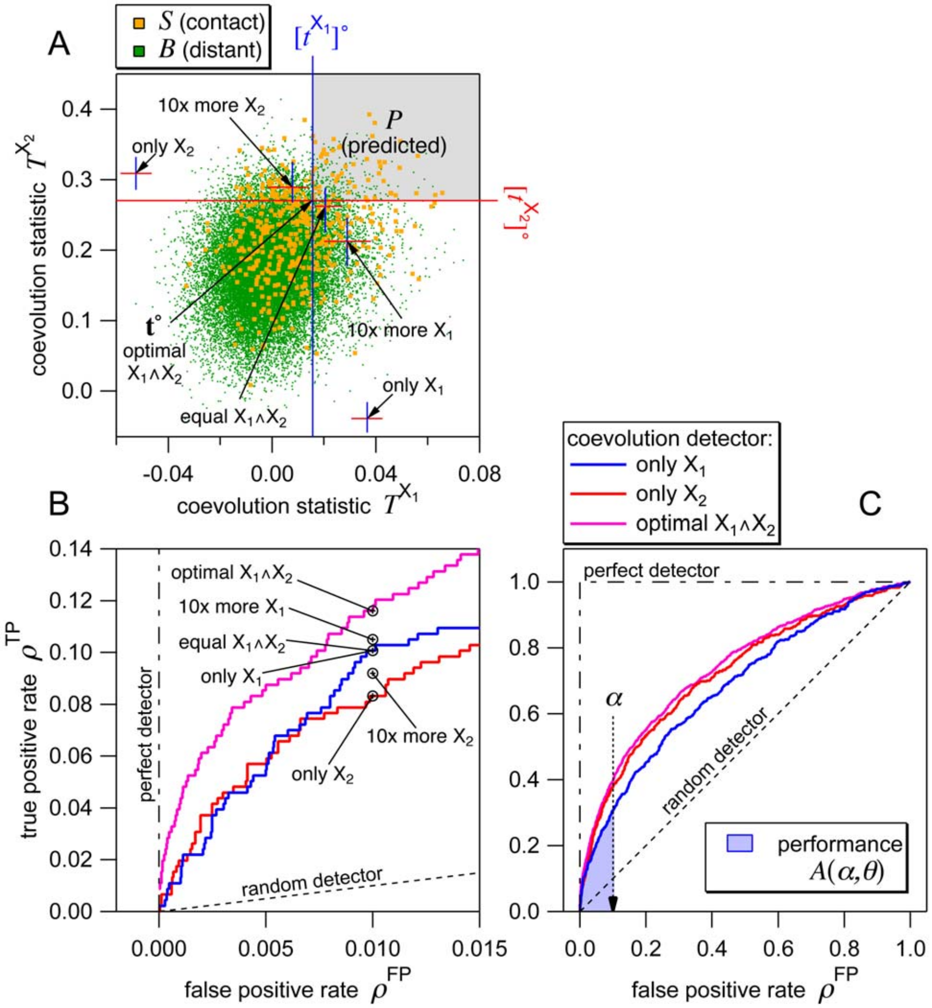

Figure 3. Weighted combination of coevolution detector X_1 and X_2. (**A**) Green and orange dots represent a set E of pairs (i,j) of amino acid positions in a protein family. $E = S \bigcup B$, where S is the set of structural contact pairs (orange) and B is the set of structurally distant pairs (green). X_1 and X_2 are coevolution detectors with statistics T^{X_1} and T^{X_2}, respectively, which are evaluated separately for each pair. A combined detector $X_1 \wedge X_2$ uses a pair of thresholds $\mathbf{t} \equiv (t^{X_1}, t^{X_2})$ to define the set of predicted pairs P (eq. 15). The set of true positives is defined as $P \bigcap S$; the true positive rate ρ^{TP} is linearly related to the number of true positives. False positives and the false positive rate ρ^{FP} are defined analogously but with B instead of S (eq. 16–17). Even if ρ^{FP} is fixed, **t** (and thus P) can still vary if t^{X_1} and t^{X_2} change in the opposite direction. Changing **t** at fixed ρ^{FP} is called *detector weighting*. For example, $\rho^{\mathrm{FP}} = 0.01$ for all 6 thresholds **t** marked by the arrowheads. For the threshold labeled as "equal $X_1 \wedge X_2$" the two detectors are combined in equal weights. "$10\times$ more X_1" refers to the weight of X_1 relative to X_2. "Only X_1" means that X_2 has zero weight and therefore $X_1 \wedge X_2$ is the same as using X_1 only. "$10\times$ more X_2" and "only X_2" have analogous meanings. Finally, the threshold denoted as $\mathbf{t}^\circ$ characterizes the optimally weighted $X_1 \wedge X_2$, which by definition has the highest ρ^{TP} for each ρ^{FP}. Black circles in (**B**) indicate ρ^{TP} for all 6 thresholds, at $\rho^{\mathrm{FP}} = 0.01$, and thus report on the corresponding performance. The optimal $X_1 \wedge X_2$ clearly outperforms the equally weighted one, which in this case happens to perform precisely as well as "only X_1 (their circles overlap). (**B**-**C**) Obtaining ρ^{TP} for all $\rho^{\mathrm{FP}} \in [0,1]$ results in receiver operating characteristic curves, which describe the performance of coevolution detectors with respect to theoretical random, and perfect, detectors. Each curve is determined by the parameter set θ, which includes **t** and therefore the weights on combined detectors. Integrating a curve on $[0,\alpha]$ yields the area $A(\alpha,\theta)$, which is used as a scalar measure of performance (eq. 18, Figure 4, 5). Conditions: $E = C_{[3,3]}$; $X_1 = \mathrm{MIp}$; $X_2 = \mathrm{CoMap}$; protein family = ABC-C; optimal phylogenetic filtering.
doi:10.1371/journal.pone.0036546.g003

$$A(\alpha,\theta) = \int_0^\alpha \rho^{\mathrm{TP}}(\gamma,\theta)\mathrm{d}\rho^{\mathrm{FP}}(\gamma,\theta), \tag{18}$$

Thus $A(\alpha,\theta)$ provides a scalar measure of performance at fixed θ and α. The interval $[0,\alpha]$ restricts ρ^{FP} below a chosen $\alpha \leq 1$. Small α is desired when high specificity (obtaining low ρ^{FP}) is more important than high sensitivity (achieving high ρ^{TP}), as in the case of this study. Note that $A(\alpha) = \alpha^2/2$ for a random detector.

Let ϕ be a relation transforming γ to α such that $\alpha = \phi(\gamma)$. In the new framework, the optimal parameter set θ^* is defined as

$$\theta^* = \arg\max_\theta A(\phi(\gamma),\theta), \tag{19}$$

replacing the initial formulation of the optimization problem (eq. 5). Thus, for each $\gamma \in [0,1]$, a unique θ^* is obtained, which is precisely the central goal of this work (eq. 6).

In the present analysis of ABC transporters $N = 11$ detectors $X_n,(n = 1,\ldots,N)$ were employed, and $K = 10$ substitution rate

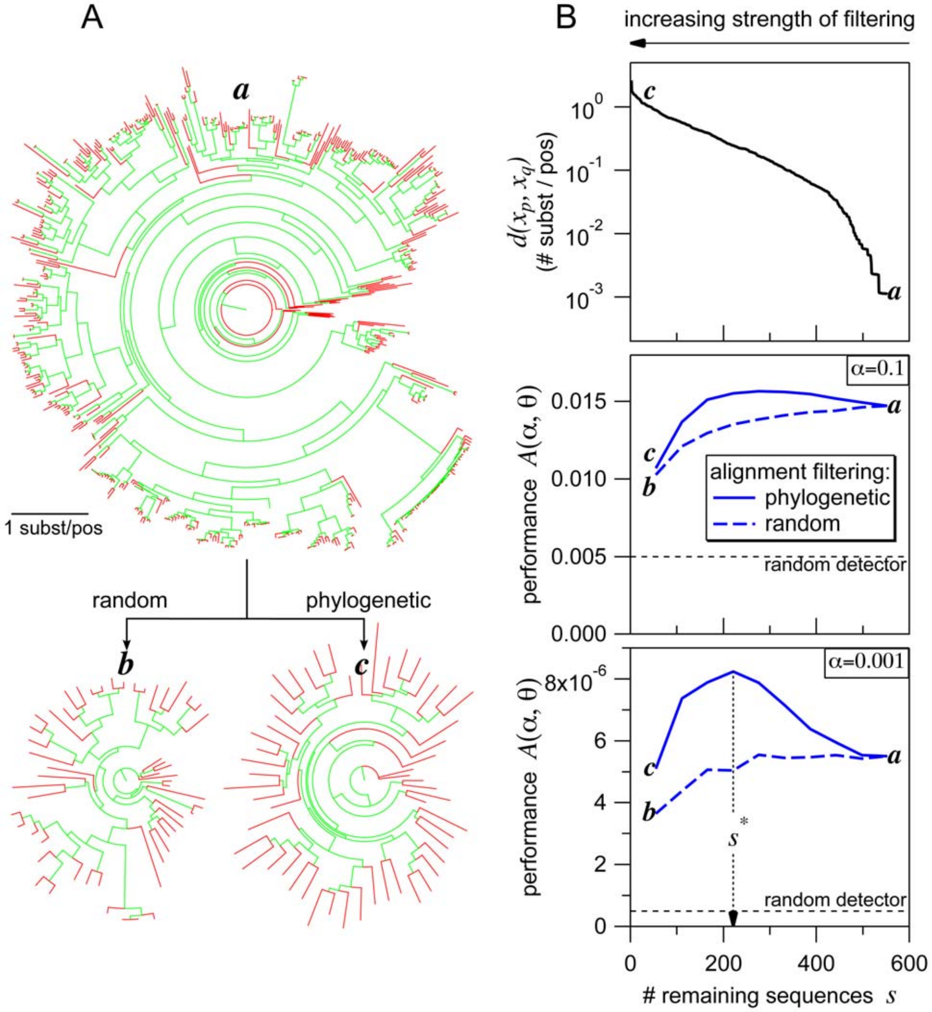

Figure 4. Influence of alignment filtering. (**A**) Random filtering and phylogenetic filtering both remove sequences from the unfiltered alignment, which is represented by the large tree ***a***, but result in trees (***b*** and ***c***) that differ in the length of terminal branches (red). Tree ***b*** (random filter) is similar to ***a*** in containing many extremely short terminal branches that are known to challenge coevolution detectors. In contrast, tree ***c*** (phylogenetic filter) lacks short terminal branches. (**B**) Opposing effects of progressively *increasing* strength of filtering, which leaves gradually *fewer* sequences in the alignment. The top graph shows, for the phylogenetic filter, the minimal sequence-sequence distance $d(x_p,x_q)$ among all sequence pairs in the filtered alignment. The two lower graphs show performance, measured by $A(\alpha,\theta)$, of a coevolution detector for both the phylogenetic and random filter. The first effect, specific to the phylogenetic filter, is a rise of $d(x_p,x_q)$ with *increasing* strength of filtering (*decreasing* number remaining sequences). This reflects the disappearance of short terminal branches, which in turn improves performance, until a maximum is reached around 250 sequences remaining. The second effect is the deterioration of performance with increasing strength of filtering, since fewer sequences provide less information for the coevolution detector. This effect is clearly seen for the random filter regardless of the number of remaining sequences but it becomes apparent for the phylogenetic filter only with strong filtering. Conditions: detector = MIp; protein family = ABC-C. Trees were plotted using FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).
doi:10.1371/journal.pone.0036546.g004

classes $C_k,(k=1,\ldots,K)$ were used. This gave $NK-1=109$ adjustable parameters $t_k^{X_n}$ under the constraint expressed by eq. 13. In addition to this, filtering at separate $s_k^{X_n}$ for each C_k and X_n provided $NK=110$ parameters and so the parameter space Θ had a dimension of $\dim\Theta=219$. (Note that in Figure 2A the same s_k is used for all X_n.) To reduce $\dim\Theta$, the present work employed a heuristic optimization strategy for eq. 19, whose details are described in Text S1 (see also Figure 3, S1 and S8).

Structural Models and Contact Pairs

The set S of contact pairs was defined as those pairs $p=(i,j)$ for which the distance $d(i,j)$ separating the C_β atom of position i from that of j is less than 8Å in a structure representing the whole protein family. The set B of distant pairs was defined by requiring $d(i,j)>30$Å. The remaining "intermediate" pairs ($8\text{A}\leq d(i,j)\leq 30$Å) were excluded from D as in ref. [37] because a large fraction of them may be connected by chains of coevolving contact pairs [40,42]. Thus θ^* was obtained using only S and D. These sets were derived separately from Sav1866 (PDB: 2HYD) [44] and CFTR (homology model [45]) representing the ABC-B and the ABC-C family, respectively.

θ^* includes the collection τ^* of optimized thresholds, which determines the set Pm of predicted pairs (eq. 15). Next, a collection $\{P\cap S_n\}$ of sets of predicted contact pairs was obtained by using $\{S_n\}$, which was derived from a set of structures that correspond to distinct conformations of the same protein. For the ABC-B family, this set contained Pgp in the inward (3G5U [46]) and outward-facing [47] conformation, and for the ABC-C family,

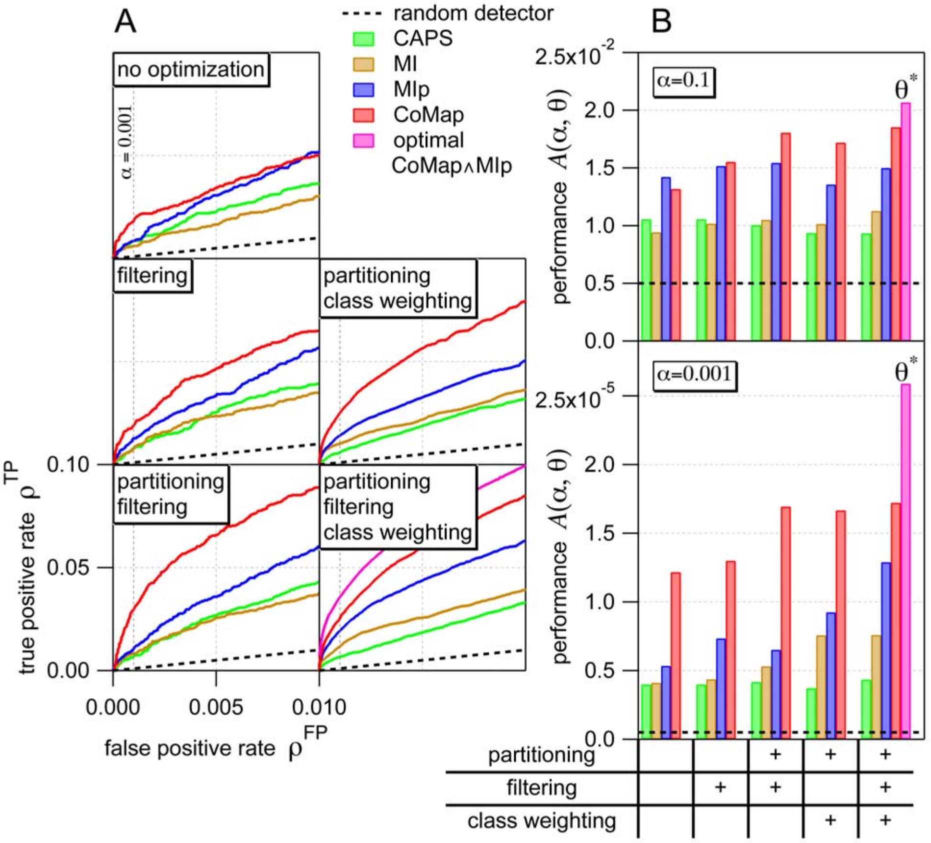

Figure 5. Optimizing the prediction of coevolving position pairs. Performance of several coevolution detectors (identified by color keys) characterized by (**A**) receiver operating characteristic curves and (**B**) partial area A under these curves. Top graph in (**B**): low specificity ($\alpha = 0.1$); bottom graph: high specificity ($\alpha = 0.001$). θ^* (above magenta bars) indicates the optimally weighted detector combination CoMap∧MIp after partitioning, optimal filtering and optimal class weighting (Figure 2). These optimal conditions yield the parameter set θ^* (eq. 19), which determines the set $P(\theta^*)$ of predicted coevolving pairs, presented in Figure 6 and Table 2, 3. These results were obtained from the ABC-C dataset.
doi:10.1371/journal.pone.0036546.g005

CFTR in the inward [48] and outward-facing [45] conformation. Consequently, a small fraction of predicted pairs were contact pairs selectively in some but not other conformations: for these pairs $(i,j) \notin P \cap S_n$ but $(i,j) \in P \cap S_m$ $(n \neq m)$.

Amino Acid Substitution Model and Rate Classes

The definition of rate classes C_k requires some discussion on the amino acid substitution model used in this study. The same model also played a role in the estimation of sequence-sequence distances (which were used for alignment filtering, as explained in the next subsection), in the inference of phylogenetic trees and in the evaluation of the coevolution statistic of certain detectors. Sequence-sequence distances and trees were both estimated by maximum likelihood using RAxML v7.0.4 [49].

The substitution of amino acid residues at each position was modeled as a continuous-time Markov process with a distinct transition rate between each pair of amino acids. The transition rates used in this study were those described by the WAG-F-Γ model [50]. In this model, the transition rates are scaled by a specific factor at each position i; the scaling factor is known as the (overall) substitution rate V^i. In other words, the substitution rate is allowed to vary among positions (p.110 of ref. [51]). Note that substitution rate is inversely related to "residue conservation".

Considering all positions, the collection $\{V^i\}$ of rates is a set of independent, identically distributed random variables. The distribution is Γ-type with cumulative density function F_Γ. Given the number M of rate classes of single positions a new random variable, the *discretized substitution rate* R^i, is defined as

$$R^i = 1 + \lfloor F_\Gamma(V^i) M \rfloor, \tag{20}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. It follows directly from definition eq. 20 that R^i takes values on $\{1,\ldots,M\}$ and has discrete uniform distribution with probability mass function $\{p_k\}$ such that $p_k = 1/M$ $(k = 1,\ldots,M)$.

This uniform "prior" probability mass function $\{p_k\}$ can be updated, for each position i, to the "posterior" the maximum likelihood estimate $\{p_k^*\}^i$ when an alignment and a tree is given. In this study this was done with CoMAP v1.3.0 [19] using the tree inferred from the alignment (which corresponds to an empirical Bayes approach; see p. 114 of ref. [51]). The *estimated* discretized substitution rate r^i of position i is defined as the mode of the posterior distribution $\{p_k^*\}^i$:

$$r^i = l \Leftrightarrow p_k^* \leq p_l^*, \forall k \in \{1,\ldots,M\}. \tag{21}$$

Given r^i and r^j for each position pair $(i,j) \in \Omega$, the class $C_{[m,n]}$ of pairs is defined as

$$C_{[m,n]} = \{(i,j)|r^i = m\mathrm{and}r^j = n\} \cup \{(i,j)|r^j = m\mathrm{and}r^i = n\}, \qquad (22)$$

where $m,n \in \{1,\ldots,M\}$. By the symmetry of the right side of eq. 22, $C_{[m,n]} = C_{[n,m]}$ so it can be required that $m \leq n$. Then the number K of classes of pairs is derived from M according to $K = (M+1)M/2$. In this work $M = 4$ and so $K = 10$ (Figure S2).

The notation $C_{[m,n]}$ can be replaced by C_k using any function that maps each $[m,n]$ to a unique k. The present work uses the simpler C_k notation to refer to a rate class in general (as in eq. 10), and the $C_{[m,n]}$ form to denote a specific class (e.g. $C_{[2,4]}$). Similarly, the symbols $P_{[m,n]}$, $t_{[m,n]}$ and $s_{[m,n]}$ have the same meaning as P_k, t_k (eq. 11–12) and s_k (eq. 14), respectively.

Multiple Sequence Alignments

A set of ABC-B and a set of ABC-C protein sequences were collected from UniProt release 15.8 using HMMER3 [52]. In both the ABC-B and ABC-C family the "full transporter" is composed of two homologous "half transporters", each of which contains a TMD and an NBD arranged as TMD-NBD (the "-" means that the domains are on the same subunit). But there are important differences between the two families. In in most ABC-B proteins the two halves constitute separate subunits (domain arrangement: TMD1-NBD1 TMD2-NBD2) while in all ABC-C proteins the halves are covalently linked (TMD1-NBD1-TMD2-NBD2). Moreover, in ABC-B proteins the two halves TMD*n*-NBD*n* ($n = 1,2$) are in general identical or very similar to each other but in ABC-C proteins the halves have extremely diverged from each other. For these reasons, the ABC-B sequence set contained half transporters but the ABC-C set contained full transporters.

A separate multiple alignment (Dataset S1 and S2) was made from each set using MAFFT v6.717b [53] from which all gap-containing positions were removed while keeping the remaining positions aligned. The resulting ABC-B alignment contained 1585 sequences, the ABC-C alignment 553 sequences.

Alignment Filtering

For each unfiltered alignment D and filter type F, a sequence $\{D(F,s)\}, s = 2,\ldots,n$, of filtered alignments was generated by removing $n-s$ sequences, where n is the number of sequences in D. As mentioned above eq. 9, the type specifies the order of removal. The two types used in this work are called *phylogenetic filter* and *random filter* (Figure 4). As discussed before, the role of the phylogenetic filter employed in this work is to remove "sequence redundancies" from the alignment. In contrast, the random filter will be used to study how the performance of coevolution detectors depend on the number of aligned sequences.

In case of the random filter, the order of removal is given by a random permutation of sequences. The phylogenetic filter applies a deterministic permutation rule to the alignment $D(F_{\mathrm{phylo}},s)$ before the next sequence is removed and $D(F_{\mathrm{phylo}},s-1)$ is generated. The rule is to consider the pair-wise evolutionary distance of all sequence pairs (x_m,x_n), where $x_m \in D(F_{\mathrm{phylo}},s), x_n \in D(F_{\mathrm{phylo}},s)$ and $1 \leq m,n \leq s, m \neq n$. Next, the pair (x_p,x_q) that has the shortest distance is found. Note that this is the most redundant pair according to the distance measure. Next, either x_p or x_q is swapped with x_1 producing the new permutation. Removing the first sequence of the new permutation creates $D(F_{\mathrm{phylo}},s-1)$ and completes the cycle. Thus s is decremented by one in each iteration of the cycle.

In terms of a phylogenetic tree, a single cycle is equivalent to finding the pair of tips connected by the shortest distance and stripping away one of these tips (with its terminal branch). As this cycle is repeated, filtering becomes "stronger", the number of sequences decreases, and the minimal sequence-sequence distance $d(x_p,x_q)$ increases in the alignment (Figure 4B top graph).

To save computational time, only a subsequence of alignments $D(F,s_k), k = 1,\ldots,10$ were analyzed with coevolution detectors. For $k = 1,\ldots,9$, $\{s_k\}$ was chosen to be uniformly spaced (within rounding error) between 1 and n, whereas s_{10} was set to n corresponding to the unfiltered alignment.

Selected Coevolution Detectors

Three families of coevolution detectors were used in this study: CoMap [19,38], mutual information (MI) [54] and CAPS [55]. The CoMap family is conceptually related to detectors in ref. [11,14,37]. This family contains detectors of the form CoMap-Y-Z, where Y is either *correlation* or *compensation*; and Z is either *simple*, *Grantham*, *polarity*, *volume* or *charge* [19]. Unlike other Zs, *simple* can be combined only with *correlation* but not with *compensation*. In this work CoMap-correlation-simple is referred to as CoMap. The mutual information family contains MI [54] and MIp [31]. The CAPS family, closely related to McBASC and other detectors [13,16], consists of CAPS and CAPS-t, where "t" denotes time correction [55].

The selected detectors strikingly differ in whether, and how, they account for the non-independence of phylogenetically related sequences. CoMap accounts for this non-independence from "first principles". This detector considers the set of branches $\{B_n\}$ of a phylogenetic tree as a sample space on which, for each position i, a random variable $X_i : \{B_n\} \rightarrow \mathbb{R}^+$ is defined, whose value is the expected number of substitutions that occurred along a given branch B_n. For each pair (i,j) the statistic of CoMap is the correlation coefficient between X_i and X_j. In contrast, MIp and CAPS-t uses empirical correction formulas, whereas MI and CAPS assumes statistical independence of sequences.

Another difference among detectors is related to the transition rates of the substitution process, which is intimately related to the physico-chemical similarities between amino acids. CoMap and CAPS allows realistic, heterogeneous rates by utilizing the empirical rate matrix of the WAG-F-Γ model. MI and MIp, however, assume the same rate for all types of transition.

Unfortunately not all detectors could be applied to all alignments. The time complexity of CAPS is $\mathcal{O}(s^2)$, where s is the number of sequences in the alignment. This made alignments with $s > 400$ intractable for CAPS in the authors' implementation [55]. Due to a segmentation fault, CoMap v1.3.0 [19] failed to run on alignments with roughly $s > 500$ and with many variable positions. For these reasons only MI and MIp were applied to the large ($s > 1500$) alignments of ABC-B sequences and a few variable positions, whose discretized substitution rate was typically $r = 4$, needed to be removed from the weakly filtered ABC-C alignments ($s \approx 500$). Consequently the size of certain rate classes, especially that of $C_{[4,4]}$, was smaller than others.

Results

The procedures of the framework described above were carried out separately for the ABC-B and ABC-C protein family. The central goal of these procedures is the optimal detection of coevolving pairs of positions, given the sequence alignment data and the structural models representing each family, as well as the selected coevolution detectors. More specifically, the procedures search for the optimal parameter set θ^* (eq. 5, 19), given a structural model and the set of contact pairs. As Figure 2A illustrates, θ in general incorporates the parameters $\{s_k\}$, which determine the strength of phylogenetic alignment filtering (eq. 9), and the parameters $\{t_k^{X_n}\}$, which control both the weights on substitution rate classes (eq. 11–13) and the weighted combination of detectors (eq. 15). Moreover, θ^* determines the set $P(\theta^*)$ of optimally predicted coevolving pairs (Figure 2B) and thus set $P(\theta^*)\bigcap S$ of pairs, which represents the coevolving subset of the known side chain contacts.

In what follows, the following questions are studied: To what extent do individual procedures improve the performance of coevolution detectors in the prediction of known contacts? What are the sources of improvement? Then, the pairs in $P(\theta^*)\bigcap S$ are further analyzed and presented in light of conformational changes.

Extent and Sources of Improvement by Optimization Procedures

Figure 5 summarizes, for the ABC-C data set, contact prediction performance under θ^* (magenta, optimal CoMap$\wedge$MIp) or under conditions lacking some or all of the optimization procedures. The receiver operating characteristic curves (Figure 5A) demonstrate that the relative performance under various conditions depends on the false positive rate ρ^{FP}, or reverse specificity. Consequently, the partial area $A(\alpha,\theta)$ under these curves reports on the relative performance in a way that depends on the upper limit α of integral of ρ^{TP} with respect to ρ^{FP} (eq. 18, Figure 5B). For most optimization procedures the relative improvement in performance was greater at high specificity ($\alpha=0.001$, bottom bar graph) than at low specificity ($\alpha=0.1$, top bar graph). Importantly, $\alpha=0.001$ is more relevant to the predicted coevolving pairs (next section) because those represent the fraction $\gamma=0.001$ of all pairs (eq. 8), whose vast majority is not in contact (the structural model contained $63\times$ more distant pairs than contact pairs).

Figure 5 also demonstrates that all optimization procedures contributed to the improved performance under θ^*. At $\alpha=0.001$, the greatest improvement was effected by the optimally weighted combination of CoMap and MIp, relative to using either of the two detectors alone. For computational efficiency (Text S1) the remaining 9 detectors were omitted from the weighted combination. Discarding these detectors may be justified by the result that they were clearly inferior to CoMap and MIp in performance (Figure 5 and Figure S5 and S6). At low ρ^{FP} (Figure 5A) and at $\alpha=0.001$ (Figure 5B) CoMap greatly outperformed even MIp. Despite this, the optimally weighted CoMap$\wedge$MIp performed markedly better than CoMap alone, which demonstrates the utility of weighted combination of detectors.

Figure 3 illustrates the principle of weighted combination of coevolution detector X_1 and X_2, and presents performance for different relative weights. The figure takes as an example $X_1=$ MIp and $X_2=$ CoMap applied to substitution rate class $C_{[3,3]}$ for the ABC-C family and demonstrates that equal weighting is not in general optimal. In this case, the equally weighted $X_1\wedge X_2$ failed to induce any improvement in performance (circles in Figure 3B) in comparison with using X_1 only. This result highlights the significance of (possibly unequal) detector weighting. As mentioned before, these effect were greater at low ρ^{FP} (compare Figure 3B to C).

To understand why phylogenetic filtering improved performance (Figure 5), it is useful to recall that this filter type was designed to remove the redundancies induced by closely related sequences, since these redundancies compromise the performance of all coevolution detectors. Figure 4 exemplifies the effects of alignment filtering for MIp; similar results were found for all other detectors (Figure S7 and S8). Comparing tree ***c*** to ***a*** in Figure 4A shows that strong phylogenetic filtering had a dual effect on the tree representing the alignment: (i) very short terminal branches (which indicate redundancies) disappeared but (ii) relatively few sequences remained in the alignment. The inverse relationship between effect (i) and (ii) was further established by applying the phylogenetic filter at gradually increasing strength (Figure 4B top).

Phylogenetic filtering had a dual effect also on performance (Figure 4B). Weak filtering (when the number remaining sequences s was between ca. 300 and 550) improved, whereas strong filtering ($s<200$) deteriorated performance. Both effects were more pronounced at $\alpha=0.001$ (bottom graph) than at $\alpha=0.1$ (middle graph).

The dual effect of the phylogenetic filter on both tree and performance suggested that the increase in performance was related to effect (i) on the tree, whereas the decrease in performance to effect (ii). This hypothesis was tested by applying the random filter, which was designed to dissect effect (ii) from (i). In line with this design, strong random filtering did not affect the distribution of the length of terminal branches (tree ***b***, Figure 4A). Performance (dashed lines in Figure 4B), however, deteriorated at increasing rate with respect to the strength of random filtering. This result, in agreement with the above hypothesis, suggests that the rate of performance deterioration by effect (ii) exceeds the rate of performance improvement by effect (i) at strong filtering. Therefore, optimizing phylogenetic filtering (by finding the maximum location s^*) is equivalent to balancing these two rates (Figure 4B, bottom).

Partitioning position pairs (explained by Figure S2) into 10 substitution rate classes C_k amplified the filtering-induced improvement in performance particularly in the case of CoMap (Figure 5). Consistently, s^* depended on C_k for all detectors, especially for CoMap (see empty circles marking s^* in Figure S8). This dependence is addressed by the combination of filtering and partitioning, which allows the conditioning of s on C_k (eq. 14).

Another benefit of partitioning was related to the possibility of weighting classes. Optimal class weighting substantially improved the performance of CoMap, MIp and MI at $\alpha=0.001$ (Figure 5). The sources of this improvement were clarified by two further results. First, the distribution of the statistic of each detector clearly depended on C_k (Figure S3 and S4). Second, the conditional version of the performance measure A was calculated given each C_k (Figure S7, S8 and in particular Figure S9). This uncovered the dependence of performance on substitution rate; the dependence was especially strong for CoMap. In light of these results, the advantage of class weighting is that it removes both types of dependence by conditioning threshold t on C_k (eq. 14).

Predicted Coevolving Pairs

When the fraction γ (eq. 8) of predicted position pairs was set to 0.001, 95 and 344 coevolving pairs were predicted for the ABC-B and ABC-C family, respectively. The roughly 4-fold difference between these numbers was due to neglecting the relatively small asymmetry between the two homologous halves of ABC-B proteins by creating an alignment from half ABC-B transporter sequences (Methods). Thus, for all pairs (i,j), both position i and j was restricted to the same half ABC-B transporter (this restriction was not used for ABC-C transporters, whose halves are greatly asymmetric).

The main focus of this study is not the entire set $P\equiv P(\theta^*)$ of predicted pairs but the subset $S\bigcap P$, where S is the set of contact pairs observed in a representative structure. For the optimization procedures, S was calculated from the outward-facing Pgp and CFTR structures for the ABC-B and ABC-C family, respectively. $S\bigcap P$ contained 41 pairs for the ABC-B and 95 pairs for the ABC-C family. For both families the positive predictive value $|P\bigcap S|/|P|$ was an order of magnitude higher than the fraction $|S|/|\Omega|$ of contact pairs in the set Ω of all pairs. For example, for the ABC-C family $|P \cap S|/|P| = 0.25$ whereas $|S|/|\Omega| = 0.011$. Consequently, the separation $j-i$ between predicted pairs (i,j) in α-helices was distributed in a way that reflected α-helical periodicity (Figure S10, Movie S1) [29,36].

As a corollary of the unequal size of the 10 substitution rate classes $\{C_{[m,n]}\}, (1 \le m \le n \le 4)$ together with the weighting of these classes, the size of sets $P_{[m,n]} \equiv P \cap C_{[m,n]}$ was also non-uniform. Most predicted pairs (i,j) fell into class $C_{[3,4]}$ (Figure S1), whose definition (eq. 22) asserts either that the discretized substitution rate r^i at position i equals 3 and $r^j = 4$ or that $r^i = 4$ and $r^j = 3$. As expected, relatively variable positions (exhibiting $r = 3$ or $r = 4$) clustered mainly in the 12 transmembrane helices (TM1-TM12), whereas relatively conserved positions ($r = 1$ or $r = 2$) were typically located in the 4 intracellular loops (ICL1-ICL4) and the two NBDs, particularly at the central dimer interface (Figure 1B). The positions from which predicted pairs were composed tended to cluster also within the TM helices (Figure 1C). The latter finding, however, does not necessarily imply a natural tendency of coevolving pairs to reside in the TM helices. Rather, it can be seen as a consequence of the previous two results that link, via substitution rate, prediction sensitivity to structural localization.

For detailed exploration of the predicted coevolving pairs (Table 1, 2, 3, Dataset S5, S6), the set $P \cap (S_{\text{out}} \cup S_{\text{in}})$ was considered, where S_{out} and S_{in} is the set of contact pairs in the outward and inward-facing conformation, respectively, of Pgp or CFTR. Thus all predicted pairs were included that were in contact in at least one of these two conformations. At the same time, d_{out}, d_{in} and

$$\Delta d = d_{\text{out}} - d_{\text{in}} \tag{23}$$

were noted, where $d_{\text{out}}(i,j)$ and $d_{\text{in}}(i,j)$ is the 3D distance separating pair (i,j) in the outward and inward-facing conformation, respectively. Therefore, Δd is the change of distance induced by the complete transition from the outward to the inward-facing conformation.

For the pairs of the ABC-B family (Table 1) and for those in the NBDs of the ABC-C family (Table 2 and Figure 6A) the set of interest was further narrowed to

$$H_1 = P \cap (S_{\text{out}} \cup S_{\text{in}}) \cap G_1, \tag{24}$$

where $G_1 = \{(i,j) | j-i > 4\}$, i.e the set of pairs fulfilling the condition that i and j are separated by more than 4 positions in the sequence. This constraint removed "obvious" contact pairs, whose distance is constrained by primary rather than secondary to quaternary structure.

For the pairs of the TMDs of ABC-C proteins (Table 3, Figure 6B and Movie S2), a more restrictive condition was used to define the set $G_2 = \{(i,j) | i \in \text{TM}m, j \in \text{TM}n, m \neq n\}$. This means that the set

$$H_2 = P \cap (S_{\text{out}} \cup S_{\text{in}}) \cap G_2 \tag{25}$$

contains those pairs (i,j) that were predicted to coevolve, for which i was observed to contact j in at least one conformation, and for which i and j localized to distinct TM helices. In this case, the notion of a "TM helix" included the helices of the ICLs since those are contiguous extensions of the *sensu stricto* TM helices. Figure 1A and 6 show that each of the 4 ICLs contains two helical extensions and a single "coupling helix" [44], and that pairs of ICLs form compact structural units that predominantly interact with a single NBD: (ICL1,ICL4) with NBD1 (Figure 6A) and (ICL2,ICL3) with NBD2. These units of 4 parallel helices are hereby termed *intracellular bundle* 1 and 2 consistently with the interacting NBD.

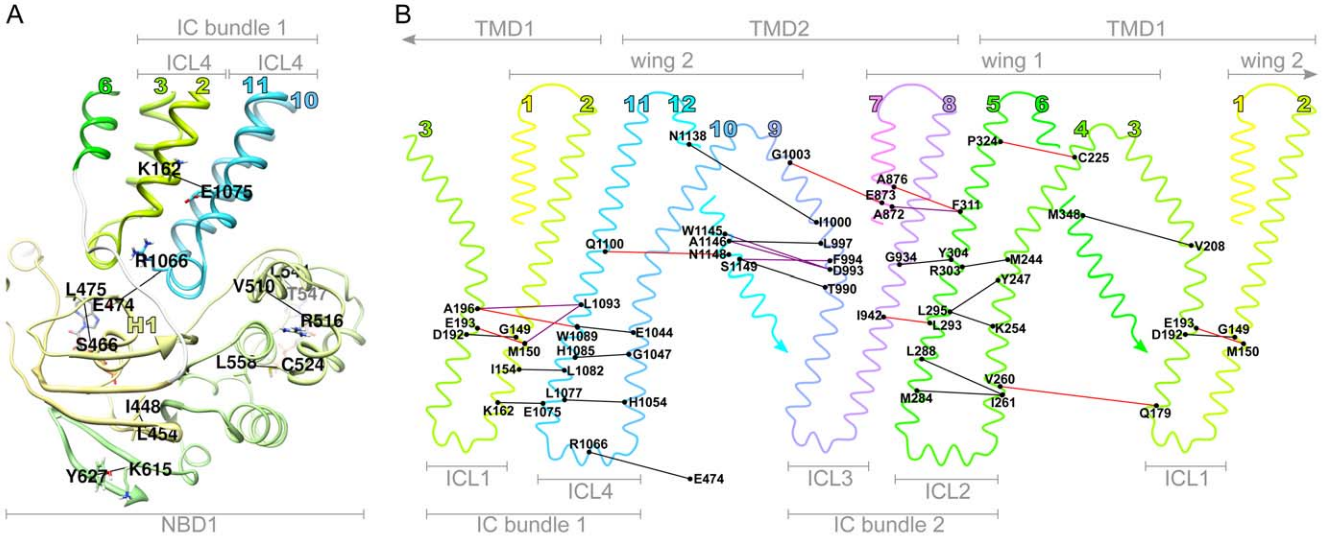

Figure 6. Coevolving position pairs in ABC-C proteins. (**A**) Labeled residue side chains connected by lines form subset H_1 of predicted coevolving position pairs (eq. 24, Table 2) in NBD1, including (E474, R1066) that connects NBD1 to ICL4. Large colored numbers identify helices of the TMDs. Helix H1 of NBD1 is also labeled as in Figure 7. (**B**) The subset H_2 of predicted pairs (eq. 25, Table 3) are indicated in a topological map of the TMD dimer, in which 12 TM helices (large colored numbers), 2 wings and 4 intracellular loops (ICLs) are labeled. The map was obtained by cylindrical projection of the two polypeptide chains of the TMD dimer. Note that TM1-TM3 are shown twice. In both **A** and **B** the color of the lines connecting predicted pairs reports on the extent of distance change $|\Delta d|$ induced by the modeled outward → inward conformational transition (eq. 23). Black: $|\Delta d| < 3$; purple: $3 \le |\Delta d| < 6$; red: $6 \le |\Delta d|$.
doi:10.1371/journal.pone.0036546.g006

Table 1. Coevolving Position Pairs in ABC-B transporters.

position i				**position** j				**3D distance (Å)**		
Pgp-N	**Pgp-C**	**region**	r^i	**Pgp-N**	**Pgp-C**	**region**	r^j	d_{out}	d_{in}	Δd
TMDs										
A58	A718	TM1	2	Q195	Q838	TM3	2	5.2	11.7	–6.4
I59	I719	TM1	3	G124	I765	TM2	3	15.5	5.4	10.1
F151	V792	ICL1	2	I369	I1012	TM6 ext.	1	7.1	11.1	–3.9
Q158	Q799	ICL1	0	N371	K1014	TMD1-NBD1	2	6.3	13.8	–7.5
S228	A871	TM4	2	A301	F944	TM5	3	5.5	9.3	–3.8
L236	L879	TM4	2	T294	I937	TM5	3	5.7	9.6	–3.9
T240	A883	ICL2	2	A361	S1004	TM6 ext.	1	7.2	27.1	–19.8
D241	L884	ICL2	3	Y363	A1006	TM6 ext.	3	6.0	30.2	–24.1
NBDs										
E393	T1036	S1	3	K416	E1059	S2	3	5.4	22.2	–16.8
R395	G1038	S1	3	M450	K1093	S4	3	5.1	21.3	–16.1
N396	E1039	S1	2	G412	G1055	S2	2	5.8	22.5	–16.7
H398	V1041	S1	3	G412	G1055	S2	2	5.3	26.0	–20.6
H398	V1041	S1	3	E448	A1091	S4	3	3.6	22.6	–19.0
S400	N1043	S1–S2 loop	3	T447	L1090	H1–S4 loop	3	4.7	21.5	–16.8
K411	Q1054	S2	3	V605	R1250	S10	3	5.5	7.5	–1.9
L415	L1058	S2	2	A599	V1244	S9	2	6.6	27.8	–21.1
Q421	Q1064	S2–S3 loop	2	V597	L1242	S9	3	6.7	5.6	1.0
V423	L1066	S3	1	V597	L1242	S9	3	5.1	3.8	1.3
V437	V1080	H1	2	L553	L1198	S7	1	5.3	15.0	–9.6
M450	K1093	S4	3	D457	E1100	S5	2	5.4	8.7	–3.2
A485	A1128	H3	3	D521	S1166	X-loop	2	6.3	17.3	–11.0
N508	N1153	H4–H4b loop	1	V568	V1213	H6	3	4.3	12.1	–7.7
V597	L1242	S9	3	K609	H1254	S10	2	6.4	19.2	–12.8

These position pairs (i,j) form subset H_1 of the predicted coevolving pairs in the ABC-B family. By definition (eq. 24), $(i,j)\in H_1$ means that i and j are in contact in either the outward or inward-facing conformation and are separated by more than four positions in the sequence. Because the ABC-B alignment contained only half transporter sequences, no pairs were predicted between the N and the C terminal halves. *Pgp-N* and *Pgp-C*: residues and positions are given for both the N and the C terminal half of human Pgp (UniProt ID: MDR1_HUMAN), respectively. The Pgp-N or Pgp-C position numbers can readily be converted to position numbers of other ABC-B half transporters using the mappings given by Dataset S3. r^i and r^j: discretized substitution rate (eq. 20) at position i and j, respectively; *3D distance*: between position i and j; d_{out} and d_{in}: distance obtained from structures representing the outward [47] and inward-facing [46] conformation, respectively; $\Delta d \equiv d_{\mathrm{out}} - d_{\mathrm{in}}$ (eq. 23). A more extensive presentation of predicted pairs is available in Dataset S5.

doi:10.1371/journal.pone.0036546.t001

Pairs Involved in Conformational Changes

Comparison of the CFTR structural models in the outward and inward-facing conformation (Movie S2) revealed possible conformational transitions [48,56]. The most striking change during the inferred outward → inward transition was the dissociation of the tight dimer of NBDs, the closure of the outward-facing cleft delineated by the wings (Figure 7A) and the opening of the inward-facing cleft between the intracellular bundles (Figure 7B). While the NBDs and the lower (i.e. proximal to the NBDs) parts of the IC bundles moved as essentially rigid bodies, the upper parts of IC bundles and especially the wings appeared flexible. A prominent component of that flexibility was the translation of some TM helices along their axes relative to other helices.

These inferred movements during the outward → inward transition were quantified by the distance change Δd (eq. 23), whose extent $|\Delta d|$ is indicated by the color of the line connecting each pair in Figure 6, 7 and Movie S1, S2, S3. In Table 2, 3, Figure 6, 7 and in the main text below residues and positions are given for human CFTR (UniProt ID: CFTR_HUMAN), whereas homologous positions for 599 other ABC-C proteins can be obtained from Dataset S4. (E873, G1003) and (Q179, V260) stood out among the pairs in H_2 (and in fact also in H_1), for which $|\Delta d|$ was relatively large (≥ 6Å, red lines). The uniqueness of these two pairs was established by the fact that they contributed to the structural contacts between the closed wings and IC bundles, respectively, but were separated by the cleft between the wings/bundles in the opposite conformation (Figure 7A-B, Movie S3).

For the rest of the red pairs (i,j) in H_1, position i resided in the same IC bundle or wing as j (Figure 6B). These included (L293, I942) in IC bundle 2, as well as (C225, P324) and (F311, A876) in wing 1. As Figure 7B and Movie 24 illustrate, the separation of these conformation-specific contact pairs was due to the inferred bending and translation of TM helix 5 with respect to TM7 and TM8. TM4 and TM5 was unusual in that they exhibited marked translation relative to each other at their extracellular ends, containing (C225, P324), whereas the same helix pair appeared relatively rigid in ICL2 (see the 4 unlabeled black and purple pairs in Figure 7B). In this regard, ICL4, formed by TM10, TM11 and a coupling helix directly interacting with NBD1, was similar to ICL2 (Figure 6B). Notably, the coupling helix of ICL4 contains

Table 2. Coevolving Position Pairs in the NBDs of ABC-C transporters.

position i				**position** j				**3D distance (Å)**		
CFTR	**region**	r^i	**ref.**	**CFTR**	**region**	r^j	**ref.**	d_{out}	d_{in}	Δd
I448	S2	3		L454	S3	3		5.1	5.1	0.0
S466	H1	1		L475	H1–S4 loop	2		7.8	7.8	0.0
V510	H3	3	[67,69]	R516	H4	3		7.3	7.2	0.1
C524	H4	2	[66]	L558	H5	1		4.8	4.9	−0.1
L541	X-loop	1		T547	C-loop	2		5.9	5.9	0.0
K615	H7–S9 loop	4		Y627	S10	3		6.8	6.8	−0.1
L1242	S3	2		I1398	S8	2		6.1	6.0	0.0
E1321	H4	3		A1391	H6	3		7.7	8.0	−0.2
K1389	H6	2		E1409	H7	2		6.4	6.2	0.1
L1399	S8	1		C1410	H7–S9 loop	2		5.8	5.7	0.1
E474	H1–S4 loop	2		R1066	coupl. H (ICL4)	1	[71–73]	7.5	9.3	−1.8

The table list those pairs (i,j) of the set H_1 (eq. 24), for which either i, j or both are located in an NBD of ABC-C proteins. For all of these pairs, except for (E474, R1066), both i and j was found in the same NBD. α-helices (H) and β-strands (S) are numbered according to ref. [74]. *CFTR*: residues and positions are given for human CFTR (UniProt ID: CFTR_HUMAN). These position numbers can readily be converted to position numbers of other ABC-C transporters using the mappings given by Dataset S4. Other columns have analogous meaning to those in Table 1 with the distinction that for this family the outward and inward-facing conformation correspond to the models described by ref. [45] and [48], respectively. A more extensive presentation of predicted pairs is available in Dataset S6.
doi:10.1371/journal.pone.0036546.t002

R1066, which together with E474 formed the only pair in H_1 that links an NBD to any other domain; $|\Delta d|$ was relatively small for this pair too.

Discussion

The new framework employed in this study is *integrative* in at least two ways. In one sense, it allows joint analysis of sequence and structural data for some protein family. In another sense, the framework integrates over several detectors by combining them in a weighted manner. In both senses, the present work surpasses previous studies, which analyzed sequence and structural data separately and used either a single detector [11,18–25] or a combined detector with equal weights [30].

How does joint analysis of sequence and structure aid the prediction of coevolving position pairs? A long-standing challenge to accurate prediction of coevolving positions has been the lack of trusted datasets on coevolution, which could help optimize the sequence-based coevolution detectors. The new framework attempts to overcome this obstacle by making use of a solved structure and defining the objective function of the optimization in terms of the prediction of known contact pairs (eq. 5, 19). The justification of this approach certainly requires some assumptions as already discussed (eq. 1–6), but these assumptions are rather weak. In particular, it is not assumed that the set of side chain contacts contain pairs that are equally tightly coupled in terms of coevolution. On the contrary: the ultimate goal of the present approach is to distinguish contact pairs that coevolve tightly from contact pairs that evolve quasi-independently. Note, however, that the new framework is inapplicable to *de novo* structure prediction problems as it relies on an existing contact map.

In its present form, the new framework takes a single input structure, representing only one conformation and only one member of the analyzed protein family. How would an alternative input structure (from the same family) influence the predictions? Although the present work does not address this question in depth, preliminary analysis indicates that switching to a different input structure affects roughly 10 to 35% of the predicted pairs depending on how different the alternative structure is relative to the original one (Figure S11). This raises the question: when multiple structures or structural models are available within a protein family, which one should be selected as structural input? Intuitively, high resolution X-ray structures are expected to be more useful inputs than lower resolution X-ray structures or homology models, and this difference might be manifested in the performance of contact prediction. Comparing a few X-ray structures and homology models in the ABC-B (Figure S12) and ABC-C (Figure S13) family indicates some differences in performance. Remarkably, performance with the 3.8 Å Pgp X-ray structure (3G5U) [46]) was lower than that with the 3.0 Å Sav1866 X-ray structure (2HYD) [44] or with the Pgp homology models [47], whose TMDs were based on the same Sav1866 structure. It remains to be determined how structural heterogeneity of homologs, as well as conformational heterogeneity within each homolog, can be accounted for to improve the prediction of coevolving residues.

Recent studies [8,9,19–21,40,42,57] presented sophisticated approaches for the prediction of higher order coevolving *networks* instead of merely coevolving *pairs*. Some of these reports [8,9,40,42] demonstrated that accounting for higher order interactions vastly improved contact prediction performance. Although the present framework ignores higher order networks, this may not undermine its power substantially because it uses contact prediction only to optimize the parameters that control coevolution detectors. It remains an open question to what extent these parameters are influenced by ignoring networks. Without doubt, the ability to infer whole networks of coevolving positions would be beneficial for the clarification of biophysical mechanisms and even for rational design of mutants, although experimental testing of ternary or higher order interactions is usually impractical (but see ref. [1]).

The new framework is quite general as it can in principle incorporate optimization procedures in addition to the three procedures used in this study: alignment filtering, class weighting and detector weighting (Figure 2A). While class and detector weighting are novel procedures, phylogenetic filtering has already been employed by the majority of published analyses of residue coevolution but with crucial differences to the current work. In all previous analyses, except ref. [22], the strength of filtering was determined by "rules of thumb", which may have lead to under or overfiltering and thus to a decline in performance, relative to even the unfiltered alignment. Moreover, it was previously ignored that the optimal filtering strength may depend on substitution rate and the selected coevolution detector, as demonstrated here (Figure S8).

Random filtering in the present work (Figure 4 and S8) revealed how performance scales with the number of sequences in the alignment [22]. The scaling itself depended both on substitution rate and the selected coevolution detector. CoMap showed the highest rate of improvement with increasing number of sequences, at least at those rates that were associated with the highest performance (Figure S8). This result suggests that CoMap can make use of the growth of sequence databases more efficiently than the other selected detectors. The same result also indicates that relatively parameter-rich, "tree-aware" detectors (like CoMap [19,38] and those in ref. [11,20,36,37]) depend more strongly on data quantity, and therefore their advantage over "tree-ignorant" detectors might have been overlooked previously [29].

Even though patterns of protein evolution may change over time, modeling time-variable patterns at the sequence level is already challenging when it is assumed that positions do not coevolve (see ref. [58] for insights). Therefore, until now, all coevolution detectors, including those in the present work, have been designed with the assumption that (co)evolutionary patterns are constant over time (i.e. persistent).

The assumption of time-invariance hinders the physico-chemical interpretation of certain pairs predicted to coevolve, while allowing time-variable patterns provides an explanation for these pairs, namely that they became coevolving from independent (or vice versa) in some lineages over time. A prime example is the pair in ABC-C proteins that corresponds to (E873, G1003) in human CFTR (Table 3 and Figure 7A), which may have become independent from coevolving as CFTR diverged away from other ABC-C proteins. Conversely, (R352, D993) was experimentally shown [59] to form a functionally important salt bridge in CFTR and yet the present analysis predicted D993 to coevolve with W1145 and A1146 rather than R352 (Table 2). But this contradiction is solved by the prediction [59] that D993 is involved in the functional divergence of CFTRs from other ABC-C proteins. For some predicted pairs, however, physico-chemical interpretation is straight-forward; e.g. (E474, R1066) in human CFTR may form a high-energy salt bridge in the solvent-inaccessible, hydrophobic interface between NBD1 and the coupling helices of two intracellular loops (Figure 6A).

Although coevolution detectors assume time-invariance, the present work did account for those changes in evolutionary patterns that occurred during long divergence processes following ancient gene duplications. As standard phylogenetic analysis suggests (Figure S14), one such duplication is the divergence of the ABC-B and ABC-C families from each other, which was followed by the divergence of the N and C terminal half transporters within the ABC-C family. These early events were taken here into account by creating separate alignment for (i.) ABC-B half transporters and (ii.) the N as well as (iii.) the C terminal ABC-C half transporters. (Note that the sequences in (ii.) and (iii.) are not separate in the sense that they form a single, "concatenated" alignment of full transporters). This approach is equivalent to ignoring the distant homology among the three clades of half transporters and has the disadvantage that those pairs cannot be identified that have persistently coevolved throughout the entire shared history of the ABC-B and ABC-C family. A related drawback is that it cannot be determined whether a predicted pair in one group of half transporters corresponds to some pair in another group, and so it cannot be studied how residue coevolution relates to the functional asymmetry between ABC-C half transporters.

All coevolution detectors use certain assumptions on the relative rates of substitution between different amino acids. The present work used CoMap with the WAG matrix [50], which derives substitution rates empirically from a large and diverse set of globular protein families. It remains to be determined to what extent this affects predictions of coevolving positions in the transmembrane domains of ABC transporters and other membrane proteins, and how the predictions would be improved by using empirical transmembrane-specific substitution matrices. The effect might be small if one considers that empirical matrices are much more similar to each other than to a "flat" matrix corresponding to unrealistic, uniform substitution rates, which is assumed by some detectors like MI.

Structural dynamics received little attention in previous coevolution analyses [8,23,37,60]. Together with a recent study [61], this report presents one of the first quantitative and systematic treatment of this question. Two classes of coevolving pairs were predicted that are distinguished by the extent $|\Delta d|$ of the 3D distance change induced by the transition between opposite-facing conformations of ABC transporters. A simple functional interpretation is that the pairs with small $|\Delta d|$ are evolutionarily conserved interactions that stabilize relatively rigid structural elements, in particular the NBDs and the intracellular bundles. In contrast, the positions of pairs with large $|\Delta d|$ appear to have coevolved with each other to stabilize selectively one (set of) conformation(s) and thus directly regulate the structural dynamics of substrate transport.

The prevalent mechanistic model of ABC transporters [3–5] emphasizes a rigid-body movement of the TMDs, which is characterized by the alternate opening and closing of the cleft between the two wings and that between the two intracellular bundles, respectively. However, only two of the predicted pairs appear to regulate the opening and closing of these clefts directly (Figure 7A). The rest of pairs with large $|\Delta d|$ (Figure 7B) were inferred to regulate relative movements of helices within the same wing or intracellular bundle. This result points toward a more refined view of conformational changes, in which TM helices bend and translate along their axes, especially in the wings, which appear to be relatively flexible.

The predicted coevolving positions in the ABC-C protein family are given here (Table 2 and 3) in terms of the sequence of human CFTR, which functions as an ion channel as opposed to all non-CFTR ABC-C proteins, which are active transporters. While this does not affect the set of predicted pairs (which can be expressed in terms of any ABC-C protein sequence using the mappings given by Dataset S4), the functional difference must be borne in mind at the mechanistic interpretation of the predictions. Since CFTR diverged away from the canonical transporter function of the family [59], it is reasonable to speculate that some fraction of coevolving pairs became uncoupled in the CFTR lineage during the divergence. Exactly what fraction of coevolving pairs has been affected depends on the extent of structural changes that conferred CFTR with its novel function, which awaits to be clarified by future structural work on CFTR. Supported by the strict coupling between ATP hydrolysis and channel gating [62], it has been hypothesized that the gating of CFTR is essentially the same as the alternating-access mechanism of an ABC-C transporter, whose internal gate has been broken by evolution [59,63]. Note that the

Table 3. Coevolving Position Pairs in the TMDs of ABC-C transporters.

(TMm,TMn)	position i				position j				3D distance (Å)		
(m,n) or (m',n')	CFTR	ICLp	r^i	ref.	CFTR	ICLq	r^j	ref.	d_{out}	d_{in}	Δd
(1,3) or (7,9)	E873		3		G1003		4		14.8	5.8	9.0
(1,11) or (7,5)	A872		3		F311		3		9.5	4.8	4.7
	A876		4		F311		3		12.7	5.4	7.3
(2,3) or (7,9)	G149	1	3	[68]	D192		3		5.3	6.4	−1.1
	M150	1	3		E193		4		13.3	6.0	7.3
(2,11) or (8,5)	M150	1	3		L1093		4		7.4	12.7	−5.4
	I154	1	3		L1082	4	3		5.7	3.7	2.0
	K162	1	3		E1075	4	4		5.7	6.7	−1.0
	G934		3		Y304		3		7.3	9.5	−2.3
	I942		3		L293	2	3		12.4	6.4	6.0
(3,4) or (9,10)	Q179	1	3		V260	2	3		5.7	16.1	−10.4
(3,6) or (9,12)	V208		3		M348		4		7.6	7.4	0.2
	T990		4		S1149		3		7.0	6.8	0.2
	D993		4	[59]	W1145		3		8.1	5.0	3.1
	D993		4	[59]	A1146		3		10.6	6.2	4.4
	F994		3		S1149		3		5.1	8.3	−3.2
	L997		4		A1146		3		5.7	7.7	−2.0
	I1000		4		N1138		3		5.6	5.4	0.2
(3,11) or (9,5)	A196		4		W1089	4	4		13.5	7.0	6.5
	A196		4		L1093		4		11.4	7.7	3.8
(4,5) or (10,11)	C225		3		P324		3		4.9	12.7	−7.7
	M244		3		R303		3		6.9	8.0	−1.2
	Y247		4		L295	2	4		7.1	7.0	0.1
	K254	2	4		L295	2	4		5.7	7.5	−1.9
	I261	2	3		M284	2	3		6.9	9.7	−2.8
	I261	2	3		L288	2	3		5.3	8.4	−3.1
	E1044	4	2		W1089	4	4		5.5	5.9	−0.3
	G1047	4	3		H1085	4	2		4.7	3.7	1.0
	H1054	4	2		L1077	4	3		7.2	8.6	−1.3
(5,6) or (11,12)	Q1100		3		N1148		2	[68]	7.9	16.1	−8.2

These position pairs (i,j) form subset H_2 of the predicted coevolving pairs in the TMDs of the ABC-C family. By definition (eq. 25), $(i,j)\in H_2$ implies that i and j are in contact in either the outward or inward-facing conformation and are located in separate TM helices. Here the notion of a "TM helix" includes the helices of the ICLs. The left column contains the indices (m,n) of each TM helix pair (TMm,TMn) together with the indices (m',n') of the homologous helix pair. *ICLp*: this column contains the index p whenever position i falls into ICLp; *ICLq* has analogous meaning for position j. For the description of all other columns see Table 1 and 2. A more extensive presentation of predicted pairs is available in Dataset S6.

doi:10.1371/journal.pone.0036546.t003

gating mechanism itself is unaffected by the regulatory (R) domain [64], another unique feature of CFTR in the ABC-C family. If the "broken gate hypothesis" holds, the extent of the function-changing structural alterations may be quite subtle, as found in the CLC channel/transporter family [65].

Recent work [26–28] showed that the combination of coevolution analysis with double mutant experiments can be a powerful tool to clarify mechanistic details of ABC proteins, although these studies focused only on a few predicted pairs in the NBDs, and in one case [26] the predicted coevolutionary coupling was not strongly supported by experimentally measured biophysical coupling. The current work offers a more complete and systematic coevolution analysis on ABC proteins. Several pairs presented here are formed by positions, at least one which was previously reported to be important for normal structure and function (see references in Table 2, 3), which hints at the practical value of the predictions. Moreover, these positions were implicated in cystic fibrosis-related folding defects of NBD1 [66], in the correction of these defects [67–69] and, as mentioned above, in CFTR channel gating [59].

This work introduces a new, integrative framework for accurate prediction of coevolving position pairs, and applies it to the ABC-B and ABC-C protein families. Each predicted pair can be interpreted as a side chain interaction that regulates some static or dynamic property of protein structure. Future experiments using site-directed mutations at these position pairs may illuminate mechanistic details that are conserved and salient features of these protein families.

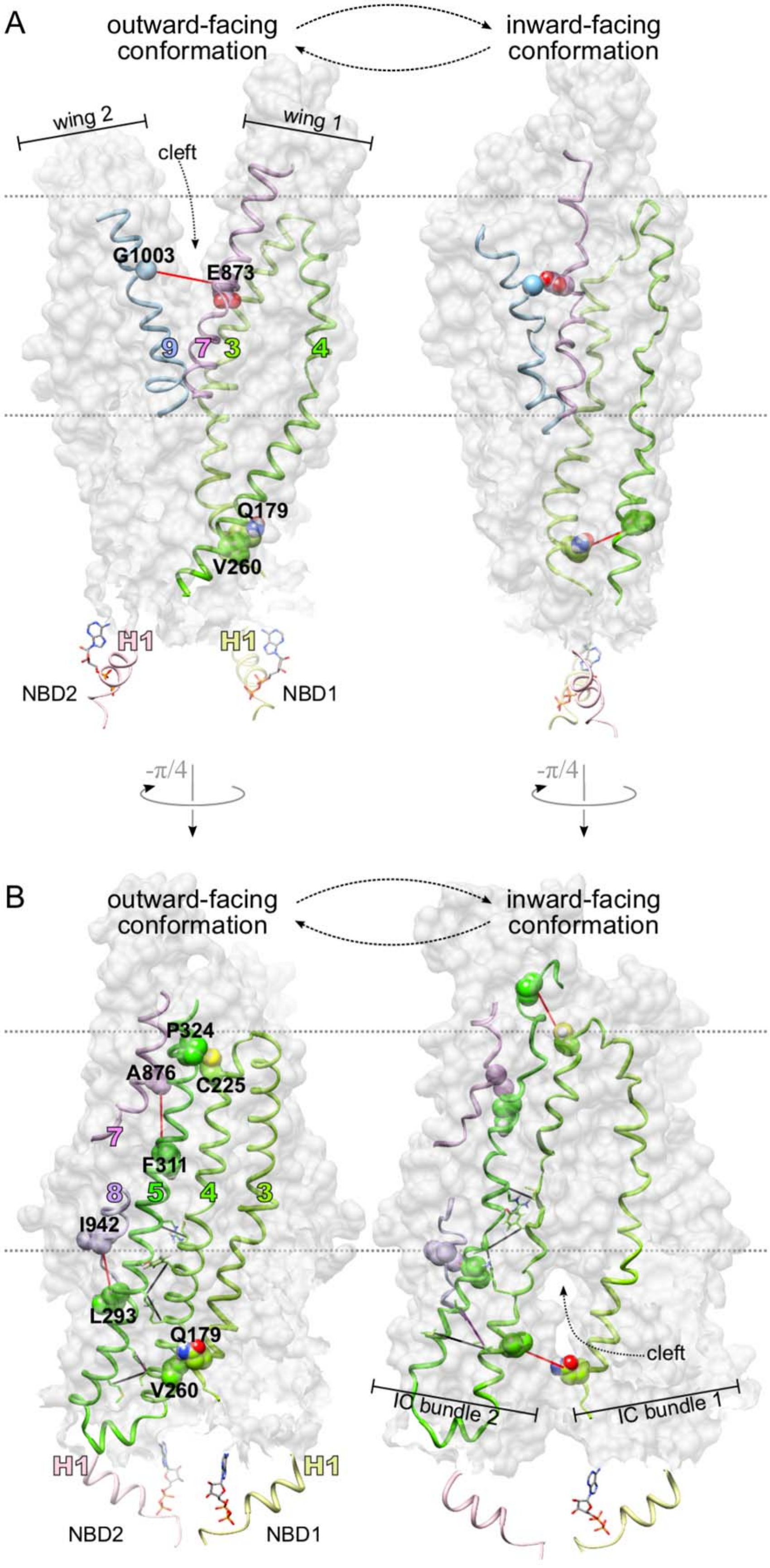
A
outward-facing conformation
inward-facing conformation
wing 2
wing 1
cleft
G1003
E873
9
7
3
4
Q179
V260
H1
H1
NBD2
NBD1
-π/4
-π/4
B
outward-facing conformation
inward-facing conformation
P324
A876
C225
7
F311
8
5
4
3
I942
L293
Q179
V260
cleft
IC bundle 2
IC bundle 1
H1
H1
NBD2
NBD1

Figure 7. Position pairs evolved to regulate conformational transitions. (**A-B**) The entire TMD dimer is shown in surface representation, and selected TM helices (identified by colored numbers) are displayed as ribbons. For each NBD, only helix H1 (cf. Figure 6A) is shown, as well as the bound ATP, if present. The outward-facing model conformation is characterized by a cleft between wing 1 and 2 (**A**, left) while the inward-facing model conformation reveals a cleft between intracellular bundle 1 and 2 (**B**, right). All labeled position pairs (connected by red lines and represented as spheres) were predicted as coevolving, and represent structural contact in only one of the two conformations, suggesting that these pairs evolved to regulate conformational transitions. Unlabeled pairs (black connecting lines, stick representation) are expected to remain in contact in both principal conformations, implying that they evolved to enhance structural rigidity. (**A**) (E873, G1003) and (Q179, V260) appear to regulate the opening and closing of the cleft between the wings and that between the IC bundles, respectively, during the conformational change. (**B**) (C225, P324), (F311, A876) and (L293, I942) might regulate the relative translation of TM4, TM5, TM7 and TM8 along the helical axes.
doi:10.1371/journal.pone.0036546.g007

Supporting Information

Figure S1 Optimization with a differential evolution algorithm. The figure shows independent runs, under various conditions defined by the control parameters of the algorithm, of the search algorithm for the optimal set τ^* of thresholds used by some coevolution detector. τ is defined as $\{t_{[m,n]}\},(1 \leq m \leq n \leq 4)$, where each $t_{[m,n]}$ is the coevolution threshold (eq. 11) corresponding to substitution rate class $C_{[m,n]}$ (eq. 22). Note that $\tau \subset \theta$ and so τ is a subset of parameters *for coevolution prediction* and is therefore not to be confused with the set of *control* parameters. The overall conclusion from this figure is that the solution τ^* identified by this heuristic algorithm is a good approximation of the global optimum. (**A**) The algorithm was run independently 12× with the same control parameters as those used for the predicted pairs presented in Table 1, 2, 3. Each run was terminated at the 1000th generation (i.e. iteration). Top graph: improvement of population fitness (defined in Algorithm 1 of Text S1) in all 12 runs. The rate of improvement declined after a few hundred generations suggesting that 1000 generations are sufficient. Bottom: the evolution of $|P_{[m,n]}(\tau)|$ is shown for one of the 12 runs (identified by black color in top graph). $P_{[m,n]}(\tau)$ is the set of predicted coevolving pairs in class $C_{[m,n]}$ and so this graph further supports the previous conclusion from the top graph. (**B**) The approximate τ^* appears to lie close to the true optimum since $\mathrm{fitness}(\tau^*) > \mathrm{fitness}(\tau_k)$, where $\{\tau_k\}$ is a random sample of size 10^6. (**C**) 1st generation (left): each of the 12 independent run was initialized from a distinct, randomly chosen, position of the parameter space. 1000th generation (right): all runs converge to nearly the same τ^*, indicated by $|P_{[m,n]}(\tau)|$. This suggests that the solution is robust against the randomness inherent to the initialization of the algorithm. (**D**) The solution appeared to be robust against also the control parameters of the algorithm.
(TIF)

Figure S2 Partitioning the set of position pairs into substitution rate classes. (**A**) Substitution rate at all 880 single positions (gray horizontal symbols) present in the ABC-C protein sequence alignment. The figure demonstrates that the substitution rate V^i varies greatly with the position index i (here the expected V^i is shown, which was obtained by the empirical Bayes approach [51], and normalized to 1 over all i). As expected (eq. 20–21), the estimated discretized substitution rate r^i (eq. 21) correlates with V^i. (**B**) Classes $C_{[m,n]}$ of pairs can be defined (eq. 22) using r^i and r^j for each of the 386760 position pairs (i,j). Since $1 \leq m \leq n \leq 4$, there are $K = 10$ classes and therefore, using a scalar index k, the partitioning results in the collection $\{C_k\}$ of classes $(k = 1, \ldots, K)$.
(EPS)

Figure S3 Dependence of coevolution statistics on substitution rate. Distribution of the standardized statistic for 4 distinct coevolution detectors (CoMap, MI, MIp and CAPS). Red line: distribution over all pairs of positions. Each blue line corresponds to the distribution over a specific rate class $C_k,(k = 1, \ldots, 10)$.
(EPS)

Figure S4 Dependence of coevolution statistics on substitution rate: tail of distribution. The graphs from Figure S3 have been expanded to illustrate the effect of substitution rate on statistical errors. Taking MIp as an example, point a marks the upper 1st percentile of the red distribution, calculated from all pairs. Setting the threshold t to the black vertical line for all pairs is equivalent to expecting the false positive rate ρ at 0.01. But since the distribution of the coevolution statistic varies substantially with substitution rate (see the dispersion of blue lines here and in Figure S3), ρ also varies at a fixed threshold. At the vertical black line, for example, ρ ranges between point c and b. Therefore the prediction is biased toward certain rate classes, such as the one identified by point b. This bias is addressed by setting a distinct threshold t_k for each class C_k (eq. 11).
(EPS)

Figure S5 Performance of variants of CoMap. The figure demonstrates that CoMap (a shorthand for CoMap-correlation-simple) outperformed other CoMap variants. These variants differ from each other in the type of coevolution statistic (correlation or compensation) and the physical quantity of the amino acid side chain that is used for the weighting of substitution vectors during the evaluation of the statistic [19]. This particular set of results corresponds to rate class $C_{[3,3]}$ but similar findings were obtained for all other classes.
(EPS)

Figure S6 Performance of variants of CAPS. The graph presents findings from a previous alignment of ABC-C protein sequences, to which a phylogenetic filter was applied. This phylogenetic filter is essentially the same as the one described in the main text and illustrated by Figure 4 except that in this case the sequence-sequence distance was expressed as (reverse) percent identity instead of the maximum likelihood estimate of the number of substitutions per position (Figure 4B top graph). In the filtered alignment the closest sequence pair had 70% identity and the time correction had essentially no effect on performance. Then a single sequence (which was previously removed by the filter) was reintroduced to the alignment. This sequence was 98% identical to some other sequence in the alignment. The bottom bar shows that time correction worsened performance to the level of a random detector. In summary, this figure demonstrates that the time correction of CAPS had either no advantage or it had an adverse effect on performance.
(EPS)

Figure S7 Random filter: performance as a function of several variables. This and the next figure explores the dependence of A on three “independent variables”: the number of remaining sequences (x axes), the substitution rate (individual graphs labeled with a particular rate class $C_{[m,n]}$) and the choice of coevolution detector (color of lines). Each solid line shows how performance scales with the number of sequences in the alignment when the distribution of sequence-sequence distance is *independent* from this number. These results correspond to the ABC-C family.
(EPS)

Figure S8 Phylogenetic filter: performance as a function of several variables. This figure is analogous to Figure S7. Each solid line shows how performance scales with the number of sequences in the alignment when the distribution of sequence-sequence distance also *depends* on this number (cf. top graph in Figure 4B). The circles indicate the optimal number s^* of remaining sequences (cf. bottom graph in Figure 4B).
(EPS)

Figure S9 Dependence of performance on substitution rate. This bubble plot shows performance, gaged by A, as the area of the circles. Performance was conditioned not only on the choice of coevolution detector (individual graphs) but also on substitution rate class (position of the circles within each graph). In principle, conditioning on rate class removes the dependence of the statistic on substitution rate (Figure S3, S4) and so dissects out the dependence of performance. Note that relative performance is displayed and that the scale at the right bottom corner depicts the area of circles that is equivalent to $1\times, 2\times$ and $4\times$ better performance than that of a random detector. The black (empty) circles represent performance at optimal phylogenetic filtering. Inside these circles gray (filled) disks represent performance without any filtering. These results correspond to the ABC-C family and should be compared to Figure S8.
(EPS)

Figure S10 Periodicity of α-helices. The histograms show the distribution of the separation $j-i$ in sequence for pairs (i,j) in the set P of predicted coevolving pairs (**A** and **C**) or in the set S of contact pairs (**B** and **D**). On the left the subset $H \cap P$ (**A**) and $H \cap S$ (**B**) is shown where H is the set of pairs (i,j) for which both i and j are located in the same helix. On the right **C** and **D** shows analogous subsets for loops instead of helices. Comparing the shapes of distributions it is clear that **A** is similar to **B**, and **C** to **D**; the resemblance is due to the high fraction of contact pairs in P. Comparing **A** to **C**, and **B** to **D** reveals a peak at $j-i=3$ or $j-i=4$ and a valley at $j-i=2$ in **A** and **B** but not in **C** and **D**. The peak corresponds to one helical turn, whereas the valley half a turn.
(EPS)

Figure S11 Effect of the input structure on the set of predicted pairs. The figure shows how the set of predicted coevolving pairs depends on the input structure. Consistency of an input structure S with the reference structure R is defined as $|P_S \cap P_R|/|P_R|$, where P_S and P_R is the set of predicted pairs using S or R as structural input, respectively. When the input and reference structure is the same ($S=R$), consistency is 1 (points at the upper left corner). But when the input and reference structures differ from each other, consistency decreases to a value that depends on the RMSD difference between the structures. Even in the "worst case" (Pgp: 3F5U) consistency is about 2/3, meaning that on average two out of three pairs predicted with the reference structure are also predicted with the alternative input structure.
(EPS)

Figure S12 Effect of the input structure on performance in the ABC-B family. This figure compares different input structures with the same detector (MIp) as opposed to Figure S8, which compares different detectors with the same input structure. Sav1866: 2HYD [44] and Pgp: closed [47] represent the outward facing conformation while Pgp: semiopen [47], Pgp: open [47] and Pgp: 3G5U [46] correspond to various inward facing conformations.
(EPS)

Figure S13 Effect of the input structure on performance in the ABC-C family. This figure is analogous to Figure S12 with ABC-C instead of ABC-B family. The input structural models were taken from ref. [45] and [48].
(EPS)

Figure S14 Divergence of half transporters during the shared history of the ABC-B and ABC-C family. This phylogenetic tree, created by the neighbor joining algorithm, shows the evolution of ABC-B and ABC-C half transporters. Although the tree is unrooted, a plausible scenario is that the common ancestor of the ABC-B and ABC-C half transporter family resides on the red branch. Following an ancient gene duplication, the two families started to diverge from each other. A subsequent duplication and gene fusion, where the red branch meets the blue branches, lead to the divergence of N and C terminal half transporters within the ABC-C family. These events created three distantly related clades of half transporters (grey shade). To avoid complications arising from functional divergence, residue coevolution was analyzed separately for each clade in the present work.
(EPS)

Text S1 Heuristic search strategy for the optimal parameter set θ^*. The text describes a stepwise strategy for obtaining an approximate θ^*. The differential evolution search algorithm of the last step is presented as pseudocode.
(PDF)

Movie S1 Predicted pairs (i,j) with separation $j-i \leq 4$ in sequence. The ribbon represents the polypeptide chain of CFTR in outward-facing conformation, and its colors match with those in Figure 6, 7 and Movie S2, S3. Residues in stick representation, connected by straight black lines, are position pairs predicted to coevolve in the ABC-C family and separated by 4 or fewer positions in sequence. For many pairs the separation occurs at one turn in an α-helix (Figure S10A). ATP molecules are shown in sphere representation.
(MOV)

Movie S2 Predicted pairs (i,j) with separation $j-i>4$ in sequence. The straight lines connect pairs contained in subset H_2 (eq. 25). As in Figure 6, black, purple and red connecting lines indicate the extent $|\Delta d|$ to which the 3D distance between i and j changes during conformational transition. The transition is modeled here by linear interpolation (morph) between the inward and outward-facing conformations.
(MOV)

Movie S3 Opening and closing of the wings and intracellular bundles of the TMDs. As Movie S2, but showing only the same two pairs (sphere representation) as Figure 7A. Note that the cleft between the wings opens as that between the intracellular bundles closes and *vice versa.*
(MOV)

Dataset S1 The ABC-B alignment. Note that all gap-containing columns have been removed.
(FA)

Dataset S2 The ABC-C alignment. Note that this alignment contains full transporters.
(FA)

Dataset S3 Positions of the ABC-B alignment. This text file is a modified version of the unfiltered alignment (Dataset S1) of ABC-C protein sequences. The modification was to substitute, for each position and sequence, the one-letter amino

acid code with the position number (position numbers are separated by commas). Therefore, this modification allows one to "translate" pairs of coevolving residue numbers in terms of Pgp (Table 1) to that in terms of any other ABC-B protein that is represented in this dataset. This is done simply by mapping residue numbers of Pgp-N (i.e. MDR1_HUMAN_N) to alignment column numbers and then column numbers to residue numbers of any protein P of interest; symbolically: position(MDR1_HUMAN_N) → column →position(P). Sequence names are given as UniProt IDs, such as MDR1_HUMAN (Pgp). "Full transporters" are represented by both of their halves: the N and the C terminal one. To distinguish between these two, the ID of the N terminal half was extended with an "_N" appendix, like MDR1_HUMAN_N. Gaps had been previously removed from this alignment, which rendered several sequences to be identical to each other, even though the corresponding full sequences were not identical. Each set of "quasi-identical" sequences gave rise to an equivalence class. In the present text file, all sequences are listed within each equivalence class. For the analysis, however, only one sequence was considered in each class while the rest was removed.
(TXT)

Dataset S4 Positions of the ABC-C alignment. This is a modified version of the ABC-C alignment (Dataset S2). See Dataset 28 for further explanation.
(TXT)

Dataset S5 List of all predicted coevolving pairs in the ABC-B family. Each Excel sheet lists the predicted coevolving pairs (including those not in structural contact) for a given fraction γ of all pairs, which determines the specificity of the prediction. Compare with Table 1.
(XLS)

Dataset S6 List of all predicted coevolving pairs in the ABC-C family. Each Excel sheet lists the predicted coevolving pairs (including those not in structural contact) for a given fraction γ of all pairs, which determines the specificity of the prediction. Compare with Table 2 and 3.
(XLS)

Acknowledgments

I thank David C. Gadsby for discussions and encouragement; Julien Dutheil for illuminating discussions and technical help with CoMap; László Csanády, András Fiser and Wladimir Labeikovsky for critical comments on the manuscript; Paola Vergani, András Váradi for discussion; Steve Lockless for discussion and help with software usage in the initial phase of this work.

Author Contributions

Conceived and designed the experiments: AGK. Performed the experiments: AGK. Analyzed the data: AGK. Contributed reagents/materials/analysis tools: AGK. Wrote the paper: AGK.

References

1. Sadovsky E, Yifrach O (2007) Principles underlying energetic coupling along an allosteric communication trajectory of a voltage-activated k+ channel. Proc Natl Acad Sci U S A 104: 19813–8.
2. Ackers GK, Smith FR (1985) Effects of site-specific amino acid modi_cation on protein interactions and biological function. Annu Rev Biochem 54: 597–629.
3. Locher KP (2009) Structure and mechanism of atp-binding cassette transporters. Philos Trans R Soc Lond B Biol Sci 364: 239–45.
4. Oldham ML, Davidson AL, Chen J (2008) Structural insights into abc transporter mechanism. Curr Opin Struct Biol 18: 726–33.
5. Higgins CF, Linton KJ (2004) The atp switch model for abc transporters. Nat Struct Mol Biol 11: 918–26.
6. Codoñer FM, Fares MA (2008) Why should we care about molecular coevolution? Evol Bioinform Online 4: 29–38.
7. Galtier N, Dutheil J (2007) Coevolution within and between genes. Genome Dyn 3: 1–12.
8. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, et al. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108(49): E1293–E1301.
9. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. (2011) Protein 3d structure computed from evolutionary sequence variation. PLoS One 6: e28766.
10. Horner DS, Pirovano W, Pesole G (2008) Correlated substitution analysis and the prediction of amino acid structural contacts. Brief Bioinform 9: 46–56.
11. Yeang CH, Haussler D (2007) Detecting coevolution in and among protein domains. PLoS Comput Biol 3: e211.
12. Shackelford G, Karplus K (2007) Contact prediction using mutual information and neural nets. Proteins 69 Suppl 8: 159–64.
13. Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18: 309–17.
14. Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7: 349–58.
15. Taylor WR, Hatrick K (1994) Compensating changes in protein multiple sequence alignments. Protein Eng 7: 341–8.
16. Neher E (1994) How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci U S A 91: 98–102.
17. Jeon J, Yang JS, Kim S (2009) Integration of evolutionary features for the identification of functionally important residues in major facilitator superfamily transporters. PLoS Comput Biol 5: e1000522.
18. Fleishman SJ, Yifrach O, Ben-Tal N (2004) An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. J Mol Biol 340: 307–18.
19. Dutheil J, Galtier N (2007) Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol Biol 7: 242.
20. Poon AFY, Lewis FI, Pond SLK, Frost SDW (2007) An evolutionary-network model reveals stratified interactions in the v3 loop of the hiv-1 envelope. PLoS Comput Biol 3: e231.
21. Carlson JM, Brumme ZL, Rousseau CM, Brumme CJ, Matthews P, et al. (2008) Phylogenetic dependency networks: inferring patterns of ctl escape and codon covariation in hiv-1 gag. PLoS Comput Biol 4: e1000225.
22. Buslje CM, Santos J, Delfino JM, Nielsen M (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics 25: 1125–31.
23. Little DY, Chen L (2009) Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. PLoS One 4: e4762.
24. Gloor GB, Tyagi G, Abrassart DM, Kingston AJ, Fernandes AD, et al. (2010) Functionally compensating coevolving positions are neither homoplasic nor conserved in clades. Mol Biol Evol 27: 1181–91.
25. Poon AFY, Swenson LC, Dong WWY, Deng W, Kosakovsky Pond SL, et al. (2010) Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of hiv-1. Mol Biol Evol 27: 819–32.
26. Szollosi A, Muallem DR, Csanády L, Vergani P (2011) Mutant cycles at cftr's non-canonical atpbinding site support little interface separation during gating. J Gen Physiol 137: 549–62.
27. Szollosi A, Vergani P, Csanády L (2010) Involvement of f1296 and n1303 of cftr in induced-fit conformational change in response to atp binding at nbd2. J Gen Physiol 136: 407–23.
28. Vergani P, Lockless SW, Nairn AC, Gadsby DC (2005) Cftr channel opening by atp-driven tight dimerization of its nucleotide-binding domains. Nature 433: 876–80.
29. Caporaso JG, Smit S, Easton BC, Hunter L, Huttley GA, et al. (2008) Detecting coevolution without phylogenetic trees? tree-ignorant metrics of coevolution perform as well as tree-aware metrics. BMC Evol Biol 8: 327.
30. Fuchs A, Martin-Galiano AJ, Kalman M, Fleishman S, Ben-Tal N, et al. (2007) Co-evolving residues in membrane proteins. Bioinformatics 23: 3312–9.
31. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333–40.
32. Tillier ERM, Lui TWH (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. Bioinformatics 19: 750–5.
33. Felsenstein J (1985) Phylogenies and the comparative method. American Naturalist 125: 1.
34. Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for coevolving residues in proteins. Bioinformatics 21: 4116–24.
35. Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 56: 211–21.

36. Pollock DD, Taylor WR, Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol 287: 187–98.
37. Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R (2005) Detecting coevolving amino acid sites using bayesian mutational mapping. Bioinformatics 21 Suppl 1: i126–35.
38. Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22: 1919–28.
39. Gouveia-Oliveira R, Pedersen AG (2007) Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. Algorithms Mol Biol 2: 12.
40. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A 106: 67–72.
41. Lee BC, Kim D (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. Bioinformatics 25: 2506–13.
42. Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6: e1000633.
43. Fawcett T (2004) ROC graphs: Notes and practical considerations for researchers. Machine Learning 31.
44. Dawson RJP, Locher KP (2006) Structure of a bacterial multidrug abc transporter. Nature 443: 180–5.
45. Mornon JP, Lehn P, Callebaut I (2008) Atomic model of human cystic fibrosis transmembrane conductance regulator: membrane-spanning domains and coupling interfaces. Cell Mol Life Sci 65: 2594–612.
46. Aller SG, Yu J, Ward A, Weng Y, Chittaboina S, et al. (2009) Structure of p-glycoprotein reveals a molecular basis for poly-specific drug binding. Science 323: 1718–22.
47. O'Mara ML, Tieleman DP (2007) P-glycoprotein models of the apo and atp-bound states based on homology with sav1866 and malk. FEBS Lett 581: 4217–22.
48. Mornon JP, Lehn P, Callebaut I (2009) Molecular models of the open and closed states of the whole human cftr protein. Cell Mol Life Sci 66: 3469–86.
49. Stamatakis A, Ludwig T, Meier H (2005) Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–63.
50. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–9.
51. Yang Z (2006) Computational molecular evolution. New York: Oxford University Press USA.
52. Eddy SR (1998) Profile hidden markov models. Bioinformatics 14: 755–63.
53. Katoh K, Kuma Ki, Toh H, Miyata T (2005) Mafft version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511–8.
54. Korber BT, Farber RM, Wolpert DH, Lapedes AS (1993) Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci U S A 90: 7176–80.
55. Fares MA, Travers SAA (2006) A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. Genetics 173: 9–23.
56. Ward A, Reyes CL, Yu J, Roth CB, Chang G (2007) Flexibility in the abc transporter msba: Alternating access with a twist. Proc Natl Acad Sci U S A 104: 19005–10.
57. Haq O, Levy RM, Morozov AV, Andrec M (2009) Pairwise and higher-order correlations among drug-resistance mutations in hiv-1 subtype b protease. BMC Bioinformatics 10 Suppl 8: S10.
58. Kosiol C, Goldman N (2011) Markovian and non-markovian protein sequence evolution: Aggregated markov process models. J Mol Biol 411(4–6): 910–23.
59. Jordan IK, Kota KC, Cui G, Thompson CH, McCarty NA (2008) Evolutionary and functional divergence between the cystic fibrosis transmembrane conductance regulator and related atp-binding cassette transporters. Proc Natl Acad Sci U S A 105: 18865–70.
60. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138: 774–86.
61. Jeon J, Nam HJ, Choi YS, Yang JS, Hwang J, et al. (2011) Molecular evolution of protein conformational changes revealed by a network of evolutionarily coupled residues. Mol Biol Evol 28: 2675–85.
62. Csanády L, Vergani P, Gadsby DC (2010) Strict coupling between cftr's catalytic cycle and gating of its cl- ion pore revealed by distributions of open channel burst durations. Proc Natl Acad Sci U S A 107: 1241–6.
63. Gadsby DC, Vergani P, Csanády L (2006) The abc protein turned chloride channel whose failure causes cystic fibrosis. Nature 440: 477–83.
64. Rich DP, Gregory RJ, Anderson MP, Manavalan P, Smith AE, et al. (1991) Effect of deleting the r domain on cftr-generated chloride channels. Science 253: 205–7.
65. Accardi A, Picollo A (2010) Clc channels and transporters: proteins with borderline personalities. Biochim Biophys Acta 1798: 1457–64.
66. Serohijos AWR, Hegedus T, Riordan JR, Dokholyan NV (2008) Diminished self-chaperoning activity of the deltaf508 mutant of cftr results in protein misfolding. PLoS Comput Biol 4: e1000008.
67. Loo TW, Bartlett MC, Clarke DM (2010) The v510d suppressor mutation stabilizes deltaf508-cftr at the cell surface. Biochemistry 49: 6352–7.
68. Pagant S, Halliday JJ, Kougentakis C, Miller EA (2010) Intragenic suppressing mutations correct the folding and intracellular traffic of misfolded mutants of yor1p, a eukaryotic drug transporter. J Biol Chem 285(47): 36304–14.
69. Wang Y, Loo TW, Bartlett MC, Clarke DM (2007) Correctors promote maturation of cystic fibrosis transmembrane conductance regulator (cftr)-processing mutants by binding to the protein. J Biol Chem 282: 33247–51.
70. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) Ucsf chimera{a visualization system for exploratory research and analysis. J Comput Chem 25: 1605–12.
71. Cotten JF, Ostedgaard LS, Carson MR, Welsh MJ (1996) Effect of cystic fibrosis-associated mutations in the fourth intracellular loop of cystic fibrosis transmembrane conductance regulator. J Biol Chem 271: 21279–84.
72. Seibert FS, Linsdell P, Loo TW, Hanrahan JW, Clarke DM, et al. (1996) Disease-associated mutations in the fourth cytoplasmic loop of cystic fibrosis transmembrane conductance regulator compromise biosynthetic processing and chloride channel activity. J Biol Chem 271: 15139–45.
73. Serohijos AWR, Hegedus T, Aleksandrov AA, He L, Cui L, et al. (2008) Phenylalanine-508 mediates a cytoplasmic-membrane domain contact in the cftr 3d structure crucial to assembly and channel function. Proc Natl Acad Sci U S A 105: 3256–61.
74. Lewis HA, Buchanan SG, Burley SK, Conners K, Dickey M, et al. (2004) Structure of nucleotidebinding domain 1 of the cystic fibrosis transmembrane conductance regulator. EMBO J 23: 282–93.