SCIENTIFIC REPORTS

OPEN

Enhancing coevolution-based contact prediction by imposing structural self-consistency of the contacts

Received: 26 March 2018
Accepted: 10 July 2018
Published online: 24 July 2018

Maher M. Kassem[1], Lars B. Christoffersen[1], Andrea Cavalli[2] & Kresten Lindorff-Larsen[1]

Based on the development of new algorithms and growth of sequence databases, it has recently become possible to build robust higher-order sequence models based on sets of aligned protein sequences. Such models have proven useful in *de novo* structure prediction, where the sequence models are used to find pairs of residues that co-vary during evolution, and hence are likely to be in spatial proximity in the native protein. The accuracy of these algorithms, however, drop dramatically when the number of sequences in the alignment is small. We have developed a method that we termed CE-YAPP (CoEvolution-YAPP), that is based on YAPP (Yet Another Peak Processor), which has been shown to solve a similar problem in NMR spectroscopy. By simultaneously performing structure prediction and contact assignment, CE-YAPP uses structural self-consistency as a filter to remove false positive contacts. Furthermore, CE-YAPP solves another problem, namely how many contacts to choose from the ordered list of covarying amino acid pairs. We show that CE-YAPP consistently improves contact prediction from multiple sequence alignments, in particular for proteins that are difficult targets. We further show that the structures determined from CE-YAPP are also in better agreement with those determined using traditional methods in structural biology.

A large and recent increase in known protein sequences has sparked an interest in using the multiple sequence alignments (MSAs) of protein families to predict native contacts in globular proteins[1], membrane proteins[2,3], as well as predicting contacts in protein-protein interfaces[4,5]. Homologous proteins from diverse organisms are likely to be similar in structure, and as a result, the sequence space explored across a protein family is highly constrained[6]. Of special interest, are pairwise coevolving amino acid positions in the MSA. These coevolution patterns have been shown to correlate strongly with spatial proximity in the native 3D structure[7].

Methods initially used to quantify the degree of pairwise positional coevolution were based on local statistical models (e.g. mutual information) that do not disentangle transitive effects often seen in proteins. An example of such a transitive effect is the observed coevolution at two amino acid positions that do not physically interact, but are both in contact with a third amino acid with which they thus covary. Current state-of-the-art methods rely on global statistical models (e.g. maximum entropy), well-known from statistical physics, to help disentangle transitive effects and, thereby, provide more robust and precise contact predictions. The maximum entropy principle is increasingly used in computational biology because of its ability to produce accurate global models given observed data (e.g. an MSA) with minimal risk of overfitting[8]. The apparently first to use the maximum entropy principle in the coevolution analysis of protein sequences was Lapedes *et al.*[9]. Similar but more recent methods are the mean field Direct Coupling Analysis approach[1,4,10] followed e.g. by pseudo-likelihood maximization[11,12], sparse inverse covariance estimation approach[13], and various machine learning methods[14–16]. Many of the methods have recently been reviewed extensively[17].

An obvious and popular use of predicted contacts is to implement them as distance restraints in protein folding simulations. The restraints can dramatically reduce the conformational search space, thereby enabling structure calculations of even large (>250 amino acid residue) proteins[1]. One major challenge is, however, that

[1]Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen, DK, 2200, Denmark. [2]Institute for Research in Biomedicine, Università della Svizzera italiana (USI), Via Vincenzo Vela 6, 6500, Bellinzona, Switzerland. Correspondence and requests for materials should be addressed to A.C. (email: andrea.cavalli@irb.usi.ch) or K.L.-L. (email: lindorff@bio.ku.dk)

the number of effective sequences (definition in Methods) needs to be sufficiently large (>~5 times the number of amino acids[18]) to ensure that enough contacts can be predicted accurately. Protein families with a sufficiently large number of sequences are, however, also more likely to have at least one experimentally solved structure, which, makes template-based modeling a more viable option[18]. A key challenge is, therefore, to decrease the required number of effective sequences to a level that enables the precise contact predictions of more protein families without experimental structures. Recently, the number of protein families, with a sufficient number of effective sequences and without homologous structures, was estimated to be ~400[18]. To increase this number and thereby decrease the required number of effective sequences, developers attempt to improve (or combine[19]) the statistical models. While there might be room for improvement, it is possible that these methods will not reach the level of precision needed to consistently compute accurate protein structures without a significant number of homologous sequences. There are, however, examples of experimentally difficult protein targets without solved structures that had enough sequences to solve the structures using coevolution[2,3,20–22].

There are two initial obstacles to overcome when using predicted contacts in structure prediction. First, one needs to decide how many contacts to include. The methods described above simply rank contacts by decreasing strength of the coevolutionary signal, but does not directly provide a natural cutoff for how strong the signal should be in order to consider a pair of residues likely to be in contact. Secondly, even with many sequences and conservative choices for how many contacts to use, one generally ends up with a number of false positive (FP) predictions, i.e. pairs of residues that show some level of coevolution, but are not in close proximity in the three-dimensional structure. In practical applications, these two problems are tightly related: One would like to include as many contacts as possible to restrain the three dimensional structure, but at the same time risk including many FPs. For example, one would on average expect ~5 of the top 20 (i.e. 25%) coevolving pairs of residues to be FPs for a 100-residue long protein with an MSA with 500 sequences, increasing to ~20 of the top 50 (40%) coevolving pairs to be FPs. For the same protein, provided only 100 sequences, one would on average expect ~8 of to top 20 (i. e. 40%) increasing to ~28 of the top 50 (55%) coevolving pairs to be FPs[18].

To circumvent the problem of noisy predictions, MacCallum and co-workers have suggested an elegant approach termed MELD[23]. The basic idea in MELD is to explicitly take into account that a fraction of the predicted contacts are wrong, and hence should not be included. In practice this is done by iteratively dividing contacts into either an 'active' or 'ignored' set, with those contacts that agree the worst with the current structural model partitioned into the ignored fraction. Thus, using the example from above, if we know that ~20 of the top 50 contacts are FPs, but not which of them, we only consider as active those 30 contacts that agree best with the structure. In this way structural self-consistency is used to guide contact assignment and structure prediction at the same time. One key limitation of this approach is that it is not always clear how large a fraction of the contacts can be ignored. Even if we know the fraction of FPs that will be present on average, it is difficult to predict this number specifically for a given protein. A different approach is to include experimental data, such as from NMR, and use self-consistency as a filter to remove false positives[24].

Here, we describe a method, called CE-YAPP, which we have developed to simultaneously determine the number of long range predicted co-evolution contacts (PCCs) and to partition these contacts into true positives (TPs) and FPs (Fig. 1). The method builds upon the automated nuclear magnetic resonance (NMR) NOESY structure determination method called YAPP (Yet Another Peak Processor)[25]. YAPP automatically assigns NOESY peaks to infer distance restraints that are subsequently used in a structure calculation. In NMR, these restraints are often the only source of information used to determine protein structures. In contrast, we designed CE-YAPP to use long-range PCCs as distance restraints and combine them with local structural information in the form of secondary structure predictions. Both YAPP and CE-YAPP share a unique protocol in which distance restraints that are in systematic violation of the protein geometry during structure calculations are automatically detected and turned off. The false-positive-detection is carried out by sampling, for each individual distance restraint, a parameter that allows us to turn off this contact with some energetic cost. These contacts/restraints are then identified as likely FPs and are discarded from the initial list of predicted contacts, thus, enhancing the contact precision by keeping the TPs.

We tested CE-YAPP using a recently-described data set, called NOUMENON, which consists of 150 MSAs and their associated crystal structures[26]. This data set has been curated to remove the selection bias seen when using protein families that have at least one experimentally solved structure. Our results show that CE-YAPP provides an effective solution to the problem of both finding a useful number of contacts and filtering FPs in a noisy prediction. In particular, we find that CE-YAPP increases prediction accuracy also when there are fewer number of sequences available. We also show that CE-YAPP can be combined with different methods for contact prediction, suggesting that the approach can be used generally to improve predictions even as methods for contact prediction continue to improve.

Results and Discussion

A framework for detecting structural self-consistency. The main goal of this work was to develop a method that enhances the precision of PCCs. We developed CE-YAPP which achieves this goal by taking an automatically chosen set of long-range (sequence-wise) PCCs and identifying the FP contacts within these. CE-YAPP performs simulations that incorporate predicted secondary structure and makes geometrical considerations of each PCC to remove those that are systematically inconsistent with the geometry (Fig. 1). More specifically, CE-YAPP begins by building an extended protein structure with fixed canonical secondary structure segments (straight α-helix or extended β-strand), based on the predicted secondary structure. The segments are structurally defined using canonical ϕ and ψ dihedral angles for the residues predicted to be α-helical or β-stranded. Subsequently, CE-YAPP performs rapid simulations, using the chosen subset of PCCs as distance restraints and allows only changes to the dihedral angles that are not fixed. A computationally-efficient energy function, that includes a van der Waals term and a restraints term, controls the structure calculation while automatically identifying systematically violated distance restraints, by sampling the weights, λ_i (see Methods).

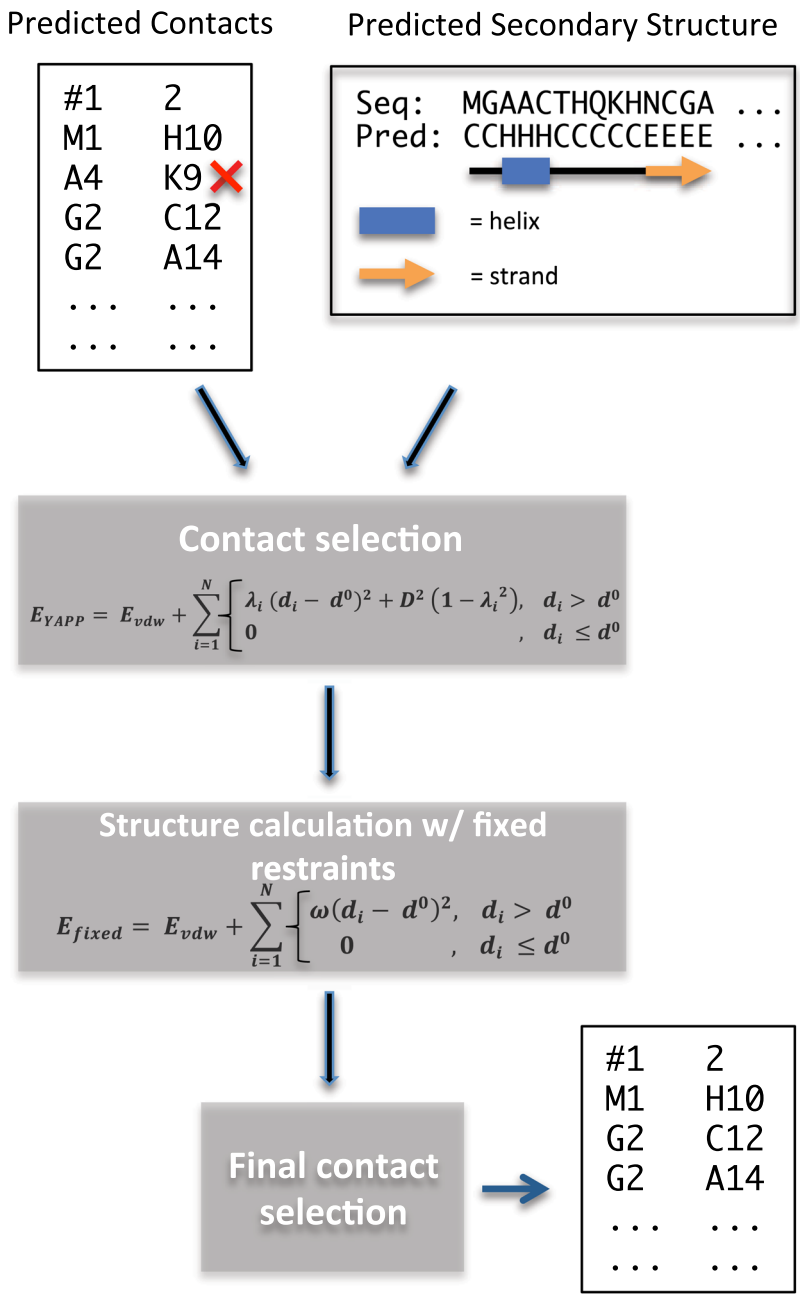

Figure 1. Workflow diagram of the CE-YAPP method. Predicted coevolution contacts and predicted secondary structure are used in combination to filter out false positive contacts. The red 'x' represents a false positive contact.

A general issue when using PCCs for structure calculation, is the need to decide the number of PCCs to use. The issue becomes especially problematic when there are only a few effective sequences (e.g. <5 sequences per N_{AA}; the number of amino acids) available, due to the higher risk of observing FP contacts[18]. In CE-YAPP we solve this issue by including a relatively large number of contacts, but then effectively filter away the FPs through requiring structural self-consistency. Specifically, we include $1.2 \times N_{AA}$ contacts between pairs of residues that are not both within a single predicted secondary structure segment (i.e. are long-range). The algorithm is robust to the exact choice of the number of contacts included (see Supporting Information for additional details).

To illustrate the idea and performance of CE-YAPP, we show the results for the 95 amino acid residues long *E. coli* ribosome hibernation promoting factor (PDB ID: 2RQL), using ~600 effectives sequences for the contact prediction (Fig. 2). In this specific case, the number of input contacts was 114 ($N_{input} = N_{AA} \times 1.2 = 114$). Thus, we first sort contacts by the strength of the evolutionary couplings and find the top 114 contacts that are not within a single predicted secondary structure element. Comparison with the known NMR structure reveals that 82 of these are TPs corresponding to a precision of 72%. To increase the precision, CE-YAPP repeats the simulation protocol (Fig. 1) 64 times and discards contacts that are turned off in more than 70% of the runs (Fig. 2a).

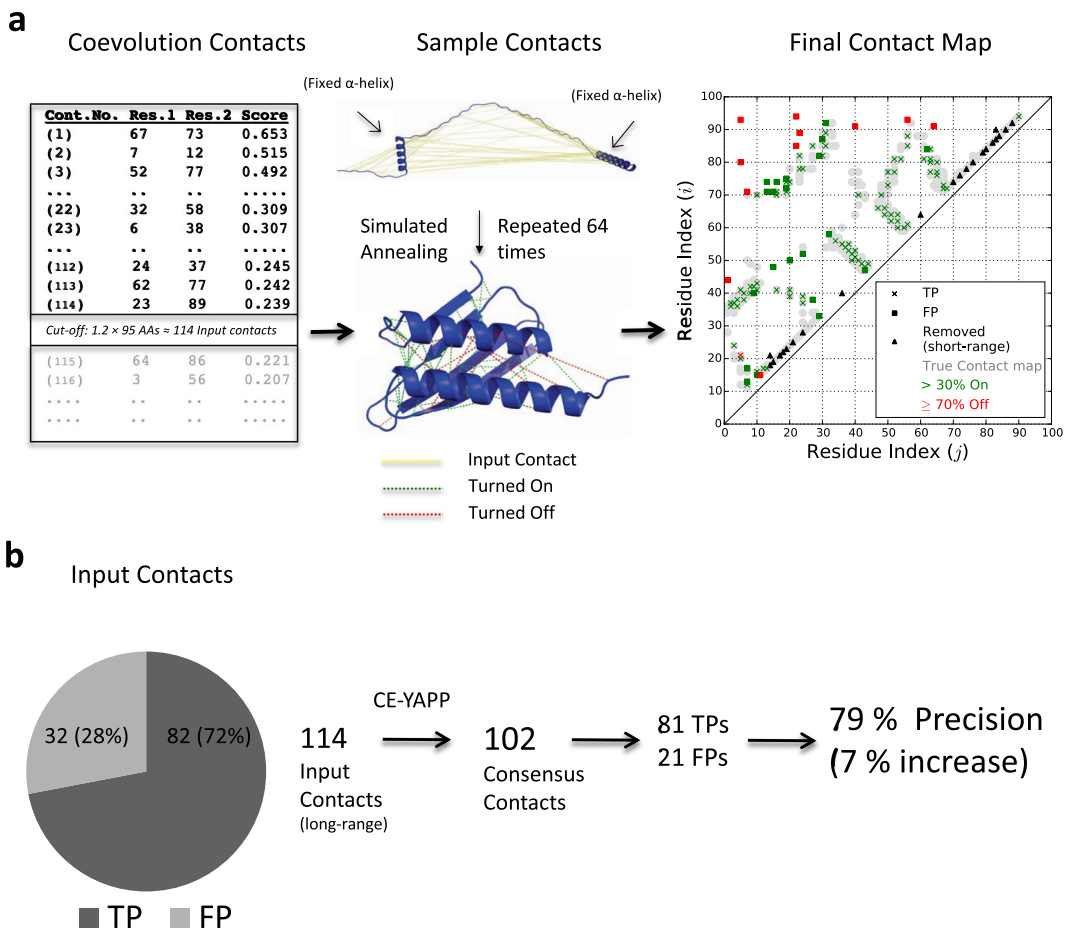

Figure 2. CE-YAPP Protocol and results for the ribosome hibernation promoting factor HPF. (**a**) CE-YAPP uses as input 114 coevolution based long-range contacts predicted using Gremlin[11]. These contacts are then used as input to the protocol depicted in Fig. 1, and repeated 64 times producing 64 similar contact lists. The final list of predicted consensus contacts are those that are turned on in more than 30% of the simulations. (**b**) The precision of the consensus contacts produced by CE-YAPP is compared to the precision of the input set of contacts.

In doing so, CE-YAPP retains 102 of the 114 contacts (CE-YAPP contacts) reducing the number of FP contacts from 32 to 21, thereby, increasing the precision from 72% to 79%. These results can be visualised in the context of the experimental contact map (Fig. 2a) which shows how most of the contacts excluded by CE-YAPP correspond to FPs, demonstrating the power of the approach in identifying a self-consistent set of contacts. The map also reveals several apparently FP contacts that are not removed by CE-YAPP. It is clear, however, that many of these are close (in sequence) to true contacts, and many of them are just outside the distance range that we use to define contacts. Thus, this case study indicates that (i) that CE-YAPP has the potential to identify a number of self-consistent contacts from a list of noisy contacts, (ii) that the algorithm can remove many FPs with only minimal loss (one contact) of TPs and (iii) it appears that at least some of the FPs that are not removed by CE-YAPP are only 'borderline errors'.

Benchmarking CE-YAPP. Encouraged by these initial observations, we continued to benchmark the performance of CE-YAPP using several indicators such as precision, recall, and number of contacts. In these analyses we used the recently described NOUMENON data set[26], which contains 150 proteins with known structures and a representative distribution of sequence depths (i.e. effective sequences); in practice we performed our analysis on 147 of these proteins (see Methods). The results, summarised in Fig. 3, demonstrate that CE-YAPP consistently improve the accuracy of contact predictions. The proteins have been sorted according to the depths of their MSAs, quantified as the number of effective sequences divided by the number of amino acids (N_{Eff}/N_{AA}; Fig. 3a). The number of input contacts (fixed at $1.2 \times N_{AA}$) and the number of contacts after running CE-YAPP are shown in Fig. 3b. As expected, when there are many effective sequences, CE-YAPP discards only few contacts whereas many are filtered away when there is only little information in the MSA. This behaviour can be rationalised given that coevolution-based contact predictors (e.g. Gremlin[11]) generally produce contacts with lower precision when N_{Eff}/N_{AA} is low, prompting CE-YAPP to discard more contacts. We also calculated the recall, i.e. the fraction of TPs in the contact list that are retained after CE-YAPP filtering (Fig. 3c). At high N_{Eff}/N_{AA} (>5), the recall

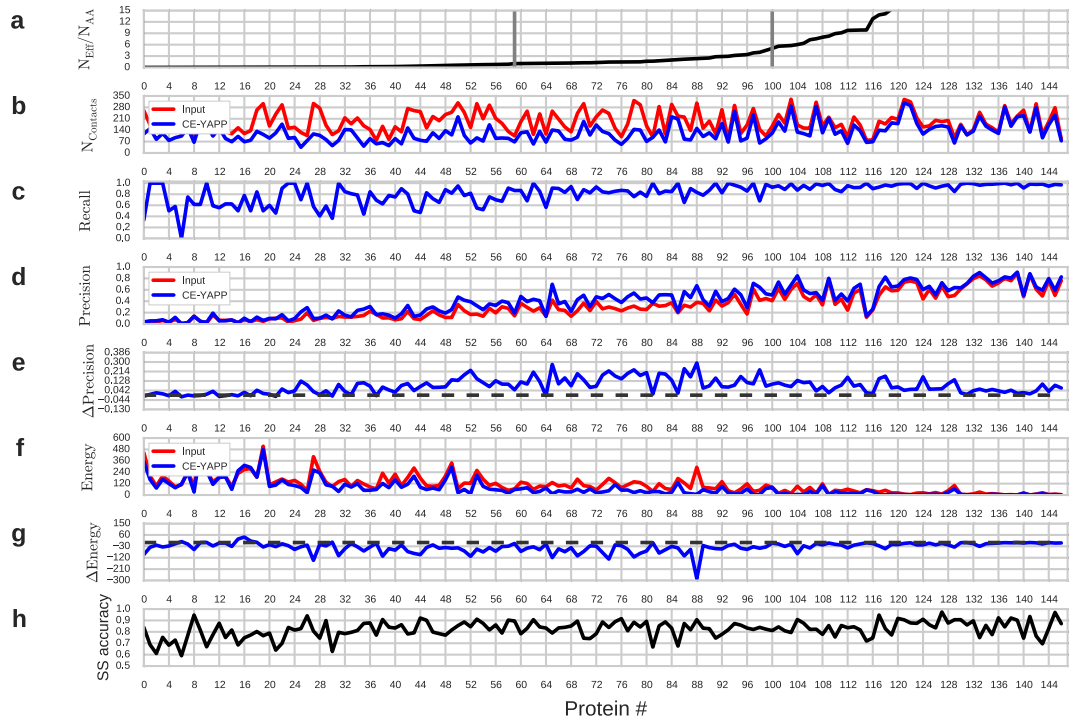

Figure 3. CE-YAPP performance on the NOUMENON dataset. (**a**) The number of effective sequences divided by the number of amino acids, N_{Eff}/N_{AA}, is plotted for each protein and sorted from low to high. The data in the remaining panels are sorted accordingly. The grey vertical bars represent the proteins with N_{Eff}/N_{AA} closest to 1 and 5, respectively. (**b**) Number of contacts. (**c**) Recall ($TP/(TP+FN)$) of the CE-YAPP contacts. (**d**) Precision ($TP/(TP+FP)$). (**e**) Precision of CE-YAPP contacts minus precision of the input contacts ($\Delta Precision$). The black dashed lines in panels (e and f) denote zero. (**f**) Restraint violation energy (Eq. 4) (**g**) Drop in restraint violation after CE-YAPP ($\Delta Energy$). (**h**) Accuracy of the predicted secondary structures using the NOUMENON multiple sequence alignments.

is close to one, meaning that CE-YAPP rarely discards TP contacts in this region. In the intermediate region ($1 < N_{Eff}/N_{AA} < 5$) the recall is ~0.8–0.9.

These results are encouraging as they suggest that CE-YAPP, even with only modest amounts of sequences, can find a consistent set of contacts that contain most of the TPs in the input set. An equally if not more important measure of performance is precision, which quantifies the fraction of contacts that are TPs. Comparison of the precision in the input contacts and the output from CE-YAPP shows a consistent improvement in precision, i.e. that CE-YAPP is able to filter away FP contacts. Again, as expected, precision is greatest at high values of N_{Eff}/N_{AA} and drops as the information content in the MSA decreases (Fig. 3d). It is clear, however, that there is a general increase in precision after CE-YAPP filtering (Fig. 3e, which shows the increase in precision after CE-YAPP). This improvement is especially pronounced with values of N_{Eff}/N_{AA} in the range 1–5 — a region that generally includes protein families that contact-based structure prediction find to be too difficult[18]. The average improvement in precision is 0.07 for $N_{Eff}/N_{AA} > 5$ and 0.14 for $1 < N_{Eff}/N_{AA} < 5$.

As discussed above in the example with the ribosome hibernation promoting factor HPF (Fig. 2) we observed that the FPs that CE-YAPP did not remove, appeared to be close to real contacts. Because precision does not quantify the severity by which FPs violate the TP definition, we also calculated the weighted mean squared distance violations ('energy'; Eq. 4) of the contacts with respect to the PDB structures (Fig. 3f), and the change of these violations after CE-YAPP (Fig. 3g). Similar to our observations using precision, we find that CE-YAPP improves contact prediction also when judged by restraint violations, and that the improvement is large also in the region with intermediate values of N_{Eff}/N_{AA} (Fig. 3g).

As expected, we note that when there are many sequences ($N_{Eff}/N_{AA} > 5$), the energy of the input contacts is significantly lower than when there are an intermediate or low number of sequences (Fig. 3f). Interestingly, this can be the case even when the apparent precision is low. Examples of this behaviour is observed for protein number 112 and 115, where the precision of the input contacts is low (~20%) but with energies close to zero. This suggests that the predicted contacts are close to the boundary between TPs and FPs, albeit more often on the 'FP side', highlighting an issue regarding precision as a performance measure.

Improved structural accuracy. Together, the results described above demonstrate how CE-YAPP can be used to find a self-consistent set of contacts, and how this algorithm is able to increase precision in contact prediction. One application of contact prediction is in protein structure prediction, where contact-assisted protein folding has enabled new progress in our ability to predict protein structure from amino acid sequence(s). Thus, we set out to examine whether the improved contact prediction also translates into improved quality of three

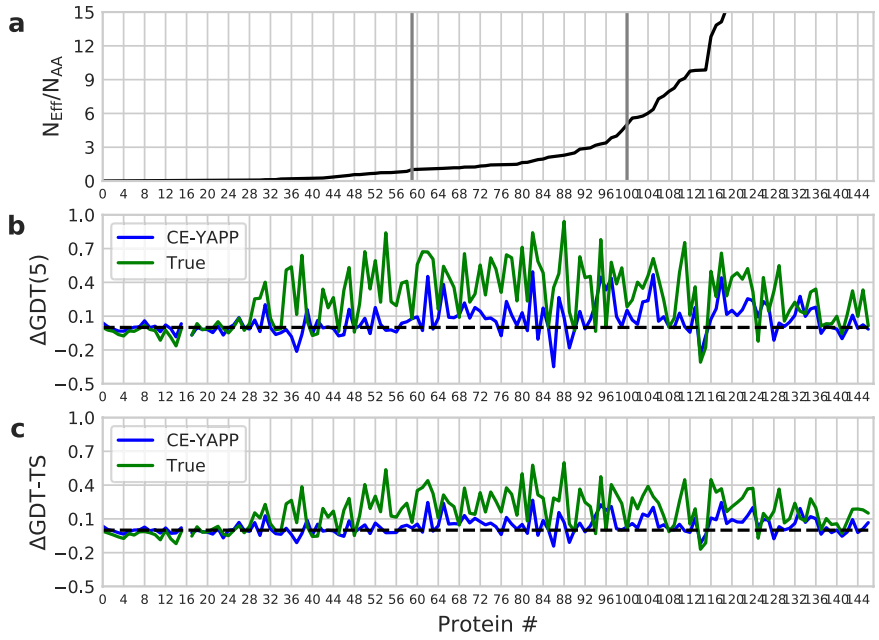

Figure 4. Structural Performance on the NOUMENON dataset. Panel (a) The number of effective sequences divided by the number of amino acids, N_{Eff}/N_{AA}, is plotted for each protein and sorted from low to high. The data in the remaining panels are sorted accordingly. The grey vertical bars represent the proteins with N_{Eff}/N_{AA} closest to 1 and 5, respectively. Panel (b) Difference in GDT(5) ($\Delta GDT(5)$). Panel (c) Difference in GDT-TS (Δ GDT-TS). The black dashed line denotes zero.

dimensional structures. In these calculations we continued to work with the long-range contacts that are the focus on CE-YAPP, but in contrast to the work described above we decided to use the actual backbone dihedral angles in the secondary structures of the experimentally-derived structures. We thus determined the ϕ and ψ dihedral angles of ordered secondary structure regions from the respective PDB structures and fixed these dihedral angles to those values. This ensures that the secondary structure in our calculations matches exactly that of the experimental structures such that we can pin down the effect of the contacts on the tertiary structure. For the same reasons, we refrained from using a complex force field to give a better picture of the contribution of the contacts to the structures, and thus used only a restraint potential and a van der Waals excluded volume term. We stress, that while we here tested the structural accuracy using dihedral angles from experimentally-derived structures, the actual contacts used in these calculations originate from running CE-YAPP using predicted secondary structures and canonical dihedral angles. As a control for the maximum performance possible, we also performed calculations using only the TP contacts from within the top-ranking contacts, where TP refers to contacts with experimental Cβ-Cβ (Cα for GLY) distances below 9 Å.

We performed 16 structure calculations for each protein and for each of the three contact sets (only TPs, before and after CE-YAPP filtering), and report the average across those repetitions, again sorting the proteins according to N_{Eff}/N_{AA} (Fig. 4). As measures of structural quality, we chose the global distance test (GDT-TS) and GDT with a single cutoff of 5 Å (GDT(5)) with high values indicating good agreement between calculated and experimental structures. Here, GDT(5) gives the fraction of Cα atoms that are within 5 Å of the experimental structure. We find that GDT(5) is a useful measure of forming the correct overall topology, given that the small distance cut-offs in GDT-TS are rarely fulfilled using our protocol. It has previously been shown that using a synthetic set consisting of only correct contacts from experimentally-derived structures, the Cα-RMSD with respect to these structures are around 2–5 Å[1]. This range of RMSDs is thus expected to be the limit for structural performance in the absence of a more detailed model or force field, also suggesting why GDT(5) is useful. We thus calculated the difference in structural accuracy when using only TPs or after CE-YAPP filtering, with respect to the structural accuracy when using input set (Fig. 4b,c). Not surprisingly, we observe that the sets of true contacts generally outperform both the input contacts and the CE-YAPP contacts with average GDT(5) values of 0.22, 0.60 and 0.73 in the three ranges of N_{Eff}/N_{AA} (<1, 1–5 and >5, respectively). The high values obtained when $N_{Eff}/N_{AA} > 1$ suggests that there are sufficiently many real contacts among the top $1.2 \times N_{AA}$ contacts to determine a reasonably accurate structure of the proteins. As described above, these calculations were performed using secondary structures defined from the experimental structure. We also performed an equivalent analysis using canonical dihedral angles based on the predicted secondary structures and, as expected, found lower values of GDT-TS and GDT(5) (Fig. S1).

In the same three ranges of N_{Eff}/N_{AA} (<1, 1–5 and >5) the average values of GDT(5) are 0.09, 0.17 and 0.48 when using contacts before CE-YAPP filtering and 0.09, 0.29 and 0.57 after filtering. From these results we make two observations. First, it is clear that although there are in principle a sufficient number of TP contacts in the middle regime to determine reasonably accurate structures, it is difficult to find these among the relatively large

number of FP contacts. Second, CE-YAPP clearly increases the structural quality also in this regime. Thus, when examining the change in GDT(5) scores ($\Delta GDT(5)$; Fig. 4c) CE-YAPP causes an average increase of 0.12 and 0.09 in the top two ranges of N_{Eff}/N_{AA}. This demonstrates that CE-YAPP is able to improve not only the contact quality but also the structural quality even when there are only an intermediate number of sequences available. Thus, for example, for the 41 proteins in the middle range we find that GDT(5) scores for 32 of the proteins are improved by CE-YAPP.

Testing other contact prediction methods. The results described above were all obtained using the Gremlin contact predictor[11] to provide the initial set of contacts to CE-YAPP. Contact prediction is, however, a field in rapid development driven both by increases in the number of sequences but also in the availability of improved algorithms[27]. These improvements are having a substantial impact on protein structure prediction, as evident from results from CASP12[28]. Because CE-YAPP is compatible with any contact predictor we analysed whether the improvements observed are specific to the use of Gremlin, or whether the requirement of structural self-consistency can generically improve a wider range of prediction methods. We thus repeated the contact predictions using four different algorithms, and used these as input to CE-YAPP. Encouragingly we observe a consistent improvement in contact predictions across all methods (Fig. S2).

Implications of predicted secondary structure accuracy and use of canonical angles. The NOUMENON database that we used to benchmark CE-YAPP is based on the observation that proteins with solved structures tend to have far more sequence homologues compared to a randomly selected protein, thus giving rise to a potential bias when judging prediction methods[26]. To avoid a similar issue when using PSIPRED to predict secondary structures, we refrained from using the default sequence database from PSIPRED and instead used only the MSA provided with NOUMENON. As expected, we find that sequences with more homologues give rise to more accurate secondary structure predictions, though with only a relatively modest dependency on N_{Eff}/N_{AA} (Fig. 3h). The trend is, however, much weaker than for e.g. the precision of the predicted contacts (Fig. 3c). While a detailed analysis is outside the scope of this work, we expect that this difference is due to the fact that contact prediction relies on creating two-dimensional histograms of sequence conservation across the entire protein sequence, whereas secondary structure prediction is fundamentally sequence-local and thus requires only accurate one-dimensional sequence profiles. These observations suggest that secondary structure prediction accuracy will generally not be the limiting factor for CE-YAPP. We do acknowledge, however, that while slightly erroneous secondary structures will likely not hamper our method, more substantial errors, such as merging of distinct secondary structural elements will increase contact prediction errors. Also, while we have achieved good results for contact prediction even when using predicted secondary structures and canonical dihedral angles, we realize that many secondary structure elements may be curved or bent. While the flexible loop regions provide enough flexibility to adapt to these differences in our structures, we do find that structural accuracy is decreased when using predicted secondary structures (Figs 4 and S1).

While we have shown that the use of predicted secondary structures and canonical dihedral angles has proven to be computationally tractable and useful in the detection of erroneous contacts, it is clear that there might be potential gains from a more accurate or flexible model of local structure. One idea is to use more flexible dihedral angle restraints based on either predictions of dihedral angles or dihedral angle distributions, as a replacement for fixed secondary structure. Moreover, one could add a physical force field (or parts of it) to more accurately model physical interactions, torsion angle distributions and hydrogen bonding. While the addition of flexible restraints or a physical force field may provide a better performance, it comes at the cost of requiring more computational power. First, by freeing up many more degrees of freedom, the structure determination step would need to sample a much larger conformational space. Second, adding a force field both requires additional computations during structure determination, and also makes the free energy landscape more rough. Both approaches will be possible in future extensions of CE-YAPP.

Equilibrium Simulations. All calculations performed above uses a simulated annealing protocol during which we both anneal the temperature (T) and the parameter (D) that determines the 'cost' of turning off a single contact (see Fig. 1 and Methods). To provide additional insight into how CE-YAPP simultaneously samples protein conformations and the λ_i-parameters that work to turn on or off each individual contact we also performed equilibrium simulations. Specifically, we simulated the 20-residue long GSGS peptide[29,30] using a Monte Carlo scheme. GSGS forms a 3-stranded anti-parallel β-sheet, and we designed a set of five contacts: two between strands β_1–β_2, two between β_2–β_3 and a fifth, erroneous contact (i.e. one that is not present in the native structure) between β_1–β_3 (Fig. S5). In our equilibrium simulations we find that the λ value for this erroneous contact rapidly decreases to zero (i.e. the contact is turned off) (Fig. S6d). In contrast, the λ values for the contacts between the neighbouring β-strands fluctuate between low and high values (Fig. S6b,c), and indeed the sum of these λ values are correlated with the RMSD to the native structure (Fig. S7), consistent with the idea that it is more energetically favourable to turn off a contact, when it is inconsistent with the geometry. For a thorough description, see the Supporting Information. In the future, it will be interesting to explore further the use of enhanced sampling methods to sample the equilibrium landscape of structure and λ parameters to understand in more detail how the CE-YAPP approach enables contact filtering by emposing structural self-consistency.

Conclusions

We have developed CE-YAPP, a method that automatically chooses a number of (long-range) predicted coevolution contacts as input and increases the precision by removing FP contacts. In its current implementation, CE-YAPP uses secondary structure prediction to define α-helical and β-stranded segments used to reduce the search space when performing efficient simulated annealing simulations. During the simulations, the weights, λ_i

(Eq. 3), are sampled to allow systematically-violated restraints to be removed and, thus, identifying them as likely FP contacts. We show, on a selection-bias-free data set consisting of 147 proteins that CE-YAPP increases the precision of PCCs. On average we observe a higher structural quality of the proteins using CE-YAPP contacts.

In the future, we expect the precision of our method should increase synergistically with the development of better contact predictors as well as the addition of system dependent experimental data such as NOEs and/or assigned chemical shift data[22,24]. We propose CE-YAPP to be used as a fast post-contact-prediction-filter before turning to more advanced structure calculations. Further, it should be possible to combine CE-YAPP with better sampling algorithms and accurate energy functions to obtain improved contact predictions and more accurate three-dimensional structures.

Methods

Simulation details. CE-YAPP uses the primary sequence, PCCs implemented as distance restraints, and predicted secondary structure of a target protein as input. Utilizing these sources of information, CE-YAPP performs structure calculations whilst simultaneously identifying and turning off distance restraints that are systematically violated by the 3D geometry. CE-YAPP then performs a final structure calculation with the refined set, keeping the distance restraints fixed. Based on this final structure, the contact list is further refined (Fig. 1). To reduce the noise levels in the refined contact list, the protocol is repeated 64 times and the consensus contacts (>30% on) are then selected as the final set of contacts.

All simulations were performed using a modified version of the YAPP method[25] implemented in the ALMOST simulation software package[31], and is available at https://sourceforge.net/projects/almost/. Simulated annealing was performed using an implementation of torsion angular dynamics[32,33] sampling only the dihedrals that are not fixed to canonical secondary structure angles based on a secondary structure prediction. An efficient energy function is used during the simulated annealing that includes a soft-core van der Waals term and a restraints term:

$$E_{YAPP} = E_{vdw} + E_{rest} \tag{1}$$

where,

$$E_{rest} = \sum_{i=1}^{N} \begin{cases} \lambda_i (d_i - d^0)^2 + D^2(1 - \lambda_i^2), & d_i > d^0 \\ 0, & d_i \le d^0. \end{cases} \tag{2}$$

Here, the sum runs over all N PCCs, d_i is the Cβ-Cβ (Cα for GLY) distance in the calculated structure, $d^0 = 7$ Å is the distance above which we consider a restraint being violated. D is a parameter used to control the acceptable degree of violation, and is decreased during the simulated annealing protocol (see below). The values of d^0, D and other key parameters were chosen as described in more detail in the Supporting Information.

During the simulations, the values of λ_i (one for each predicted contact) are updated at each time step using a Brownian motion-like equation:

$$\lambda_i(t + \Delta t) = \lambda_i(t) + \gamma F_i(t)\Delta t + \delta\sqrt{T\,\Delta t}\,\phi_i^{norm} \tag{3}$$

Here t is the MD simulation time, Δt is the time step, T is the temperature, $F_i(t) = \partial E_{rest,i}/\partial r$, is the force exerted by the restraint i at a given time t and ϕ_i^{norm} is random noise generated from a standard normal distribution. The parameters γ and δ were set to 0.00025 and 0.6666, respectively. All values of λ were enforced to stay in the range of 0–1.

By sampling λ_i during the simulations, CE-YAPP can switch specific distance restraints off ($\lambda_i = 0$) at an energetic cost determined by D^2. During the simulated annealing protocol, D is annealed from 150 Å until it reaches 3 Å to steadily remove contacts that are systematically violated. In other words, the simulation begins with all restraints turned on ($\lambda = 1$) following a subsequent annealing of D which reduces the penalty (D^2) for turning of a contact such that restraints inconsistent with the geometry will steadily be removed. A final structure calculation is performed using the refined list of contacts as fixed restraints ($\lambda_i = 1$) in a simulated annealing simulation. The restraints that violate the upper limit d^0 by more than D, in the final structure, are turned off.

To reduce noise further, we repeat the protocol (Fig. 1) 64 times producing 64 similar contact lists. The repetitions are trivially independent and can, therefore, run simultaneously on a multi-core computer. The contacts that are turned on in more than 30% of the 64 refined contact lists produced by the 64 repeated protocol runs are then selected for and represent the final contacts produced by CE-YAPP.

In the evaluation of the distance violations of the final set of contacts we also calculated the restraint violation energy:

$$E = \frac{1}{N}\sum_{i}^{N} \begin{cases} (d_i - 9)^2, & d_i > 0 \\ 0, & d_i \le 9 \end{cases} \tag{4}$$

where N is the number of contacts and d_i are the contact distances (CB-CB) observed in the PDB structures (CA for Glycine).

Fixing the secondary structure. We reduced the conformational space by fixing the secondary structure of simulated proteins to that predicted by PSIPRED[34]. The dihedral angles (ϕ and ψ) of the segments predicted to be either α-helical or β-stranded were fixed to the angles ($\phi_\alpha = -60$, $\phi_\beta = -135$ and $\psi_\alpha = -45$, $\psi_\beta = 135$) for the respective secondary-structure-type leaving only the remaining regions to change conformation. When using PSIPRED, we refrained from using the default sequence database and used only the MSAs provided with NOUMENON to minimize bias that might occur from using the larger default database when predicting secondary structures.

Effective sequences. The number of effective sequences were calculated by clustering sequences with more than 80% sequence identity and assigning each sequence within the clusters with a fractional weight of $1/n$, where n is the cluster size. By summing the weights of each sequence one obtains the effective number of sequences which represents the number of diverse sequences in the alignment.

Contact prediction. We predicted contacts using the stand-alone Gremlin[11] software package using the default settings. Using the predicted secondary structure information, we only select PCCs that do not coincide within a single predicted secondary structure segment, to probe the extraction of long-range contacts. More specifically, we optimised the number of input contacts to be 1.2 times the number of amino acids ($N_{input} = 1.2 \times N_{AA}$). We thus chose this number of contacts among those not found within a fixed secondary structure segment, and used these contacts as distance restraints.

In the analysis of the contacts we define a TP as being a predicted contact with a Cβ-Cβ (Cα for GLY) distance observed to be at or below 9 Å in the experimental structure.

Benchmarking structural accuracy. We performed simulations to determine the structural quality obtained from the different sets of contacts (Fig. 4) using the experimentally-observed dihedral angles extracted from the PDB structures. More specifically, we used STRIDE[35], to determine the secondary structure of the proteins based on the PDB structures, and we extracted the ϕ and ψ dihedral angles those residues that were determined (by STRIDE) to be α-helical, 3_{10} helical or β-stranded. During the simulation, these dihedral angles were kept fixed. In these simulations we also fixed $\lambda_i = 1$ in Eq. 3 thereby keeping the restraints fixed. We used GDT as a quality measure with a single cut-off of 5 Å. Specifically, we calculated the fraction of Cα atoms in the structural model that are within 5 Å (GDT(5)) of the corresponding position in the PDB structure. To reduce the noise levels, we take the average GDT(5) of 16 simulations for each set of contacts.

Computational time. Once the predicted secondary structure (by PSIPRED) and coevolution contacts (from Gremlin) were obtained, the time spent on a single protocol run (Fig. 1) using a single CPU-core (2.3 GHz) takes in the order of 15 CPU-minutes on any of the tested proteins. Thus, with a 64 core machine, the entire protocol can be performed in about 15 mins.

Protein data. Our benchmark of CE-YAPP was performed using the NOUMENON data set[26], which consists of 150 MSAs and their associated protein crystal structures. Three out of the 150 data points were left out of the analysis, simply because their predicted contacts all coincided in unresolved regions of the PDB structures. In particular, when there are only very few effective sequence, Gremlin may score all pairs of columns in the MSA equally, with top ranked contacts then arbitrarily assigned to the N-terminal region. For the three excluded proteins, the N-terminal tails are not resolved in the crystal structures, resulting in data points that we cannot verify against experiments.

References

1. Marks, D. S. *et al*. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
2. Hopf, T. A. *et al*. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
3. Nugent, T. & Jones, D. T. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc. Natl. Acad. Sci. USA* **109**, E1540–E1547 (2012).
4. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in proteinprotein interaction by message passing. *Proc. Natl. Acad. Sci. USA* **106**, 67–72 (2009).
5. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).
6. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Bio* **193**, 693–707 (1987).
7. Altschuh, D., Vernet, T., Berti, P., Moras, D. & Nagai, K. Coordinated amino acid changes in homologous protein families. *Protein Eng.* **2**, 193–199 (1988).
8. Boomsma, W., Ferkinghoff-Borg, J. & Lindorff-Larsen, K. Combining Experiments and Simulations Using the Maximum Entropy Principle. *PLoS Comput. Biol.* **10**, e1003406 (2014).
9. Lapedes, A., Giraud, B. & Jarzynski, C. Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. Preprint at https://arxiv.org/abs/1712.06527 (2012).
10. Morcos, F. *et al*. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* **108**, E1293–E1301 (2011).
11. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S.-I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins: Struct., Funct., Bioinf.* **79**, 1061–1078 (2011).
12. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E – Stat. Nonlinear, Soft Matter Phys.* **87**, 012707 (2013).
13. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinforma.* **28**, 184–190 (2012).
14. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultradeep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).
15. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture mutation effects. Preprint at https://arxiv.org/abs/1712.06527 (2017).
16. Figliuzzi, M., Barrat-Charlaix, P. & Weigt, M. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35**, 1018–1027 (2018).
17. Oliveira, S. H. P., Shi, J. & Deane, C. M. Comparing co-evolution methods and their application to template-free protein structure prediction. *Bioinforma.* **33**, 373–381 (2017).
18. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. USA* **110**, 15674–15679 (2013).
19. Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinforma.* **31**, 999–1006 (2015).

20. Tian, P. *et al*. Structure of a Functional Amyloid Protein Subunit Computed Using Sequence Variation. *J. Am. Chem. Soc.* **137**, 22–25 (2014).
21. Ovchinnikov, S. *et al*. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife* **4**, e09248 (2015).
22. Kassem, M. M., Wang, Y., Boomsma, W. & Lindorff-Larsen, K. Structure of the Bacterial Cytoskeleton Protein Bactofilin by NMR Chemical Shifts and Sequence Variation. *Biophys. J.* **110**, 2342–2348 (2016).
23. MacCallum, J. L., Perez, A. & Dill, K. A. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. USA* **112**, 6985–6990 (2015).
24. Tang, Y. *et al*. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat. Methods* **12**, 751–754 (2015).
25. Cavalli, A. & Vendruscolo, M. Analysis of the performance of the CHESHIRE and YAPP methods at CASDNMR round 3. *J. Biomol. NMR* **62**, 503–509 (2015).
26. Orlando, G., Raimondi, D. & Vranken, W. F. Observation selection bias in contact prediction and its implications for structural bioinformatics. *Sci. Reports* **6**, 36679 (2016).
27. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Reports on Prog. Phys.* **81**, 032601 (2018).
28. Schaarschmidt, J., Monastyrskyyn, B., Kryshtafovych, A. & Bonvi, A. M. Assessment of contact predictions in casp12 co-evolution and deep learning coming of age. Proteins: Struct. Funct. Bioinforma. **86**, 51–66 (2018).
29. Ferrara, P. & Caflisch, A. Folding simulations of a three-stranded antiparallel beta-sheet peptide. *Proc. Natl. Acad. Sci. USA* **97**, 10780–10785 (2000).
30. Ferrara, P. & Caflisch, A. Native topology or specific interactions: what is more important for protein folding? *J. Mol. Biol.* **306**, 837–850 (2001).
31. Fu, B. *et al*. ALMOST: An all atom molecular simulation toolkit for protein structure determination. *J. Comput. Chem.* **35**, 1101–1105 (2014).
32. Güntert, P., Mumenthaler, C. & Wüthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Bio* **273**, 283–98 (1997).
33. Jain, A., Vaidehi, N. & Rodriguez, G. A fast recursive algorithm for molecular dynamics simulation. *J. Comput. Phys.* **106**, 258–268 (1993).
34. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, W349–W357 (2013).
35. Frishman, D. & Argos, P. Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Struct. Funct. Genet.* **23**, 566–579 (1995).

Acknowledgements

The authors thank Wouter Boomsma for useful discussions and input.

Author Contributions

A.C. modified the YAPP protocol to function with coevolution contacts. M.M.K. automated the protocol and performed main simulations. L.B.C. performed the benchmarking simulations and analysis on other contact prediction methods. M.M.K. performed the main analysis and wrote the manuscript with the input and guidance of K.L.-L. and A.C. Both A.C. and K.L.-L. supervised and designed the project.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-29357-y.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.