BIOINFORMATICS

Vol. 19 no. 16 2003, pages 2039–2045
DOI: 10.1093/bioinformatics/btg278

Inferring protein interactions from phylogenetic distance matrices

Jason Gertz[1], Georgiy Elfond[2], Anna Shustrova[2], Matt Weisinger[3], Matteo Pellegrini[4,*], Shawn Cokus[5] and Bruce Rothschild[5]

[1]Department of Mathematics, 310 Malott Hall, Cornell University, Ithaca, NY 14853-4201, USA, [2]Department of Mathematics, 970 Evans Hall #3840, University of California, Berkeley, CA 94720-3840, USA, [3]Division of Engineering and Applied Sciences, Pierce Hall, 29 Oxford Street, Cambridge, MA 02138, USA, [4]Protein Pathways, Woodland Hills, CA, USA and [5]Department of Mathematics, University of California, Los Angeles, CA, USA

Received on September 30, 2002; revised on March 3, 2003; accepted on April 25, 2003

ABSTRACT

Finding the interacting pairs of proteins between two different protein families whose members are known to interact is an important problem in molecular biology. We developed and tested an algorithm that finds optimal matches between two families of proteins by comparing their distance matrices. A distance matrix provides a measure of the sequence similarity of proteins within a family. Since the protein sets of interest may have dozens of proteins each, the use of an efficient approximate solution is necessary. Therefore the approach we have developed consists of a *Metropolis Monte Carlo* optimization algorithm which explores the search space of possible matches between two distance matrices. We demonstrate that by using this algorithm we are able to accurately match chemokines and chemokine-receptors as well as the tgfβ family of ligands and their receptors.

Contact: matteope@proteinpathways.com

1 INTRODUCTION

Over the past few years advances in sequencing technology have allowed scientists to determine all the genes that are coded by an organism's genome. In order to characterize the function of these genes it is necessary to elucidate their interactions. Over the past few years, both experimental and computational approaches have been developed to study protein–protein interactions. Examples of experimental approaches include the yeast two hybrid technique, immunoprecipitation and tandem affinity purification tags (Uetz *et al*., 2000; Gavin *et al*., 2002). These have been extensively applied to yeast proteins to determine thousands of interactions.

Computational techniques to elucidate protein interactions include the use of phylogenetic profiles, conserved operons, protein fusions and correlated mutations to infer protein interactions (Pellegrini *et al*., 1999; Overbeek *et al*., 1999; Marcotte *et al*., 1999; Pazos and Valencia, 2002). Although these methods successfully link proteins that participate in the same cellular processes, these links represent both direct physical interactions and indirect interactions mediated by other molecules. Furthermore, these techniques often link protein families together, and do not allow one to determine which member of one family interacts with which member of the other. It is therefore useful to extend these approaches with new methodologies to probe only direct protein interactions and to find the correct matches between the members of two interacting families.

The approach we present here consists of searching for protein interactions by finding the correct matches between the leafs of two phylogenetic trees of families of interacting proteins. We assume that two interacting proteins evolve in a correlated fashion, and therefore reconstructing phylogenies allows us to infer interactions. In support of this assumption, it has previously been shown that the phylogenetic trees of various protein families show correlated evolution (Goh *et al*., 2000; Goh and Cohen, 2002). Furthermore, it has also been shown that duplicated genes in yeast tend to preserve the same sets of interactions for several million years, although eventually, after 200 million years, these patterns disappear (Wagner, 2001).

In support of the notion that gene phylogenies are useful for reconstructing interactions, we show the trees of two interacting protein families in Figure 1. These trees contain the members of the tgfβ ligands and their receptors. To illustrate qualitatively that interacting members tend to co-evolve we have colored ligands and their respective receptors in the same colors, thus illustrating the similarities between the topologies of the two trees.

The algorithm we present here identifies interacting pairs of proteins between two families, where the members of each family are related by sequence similarity. Identifying

[*]To whom correspondence should be addressed at 21111 Oxnard Blvd, Woodland Hills CA 91367, USA.

Downloaded from https://academic.oup.com/bioinformatics/article/19/16/2039/242166 by U.S. Department of Justice user on 17 August 2022

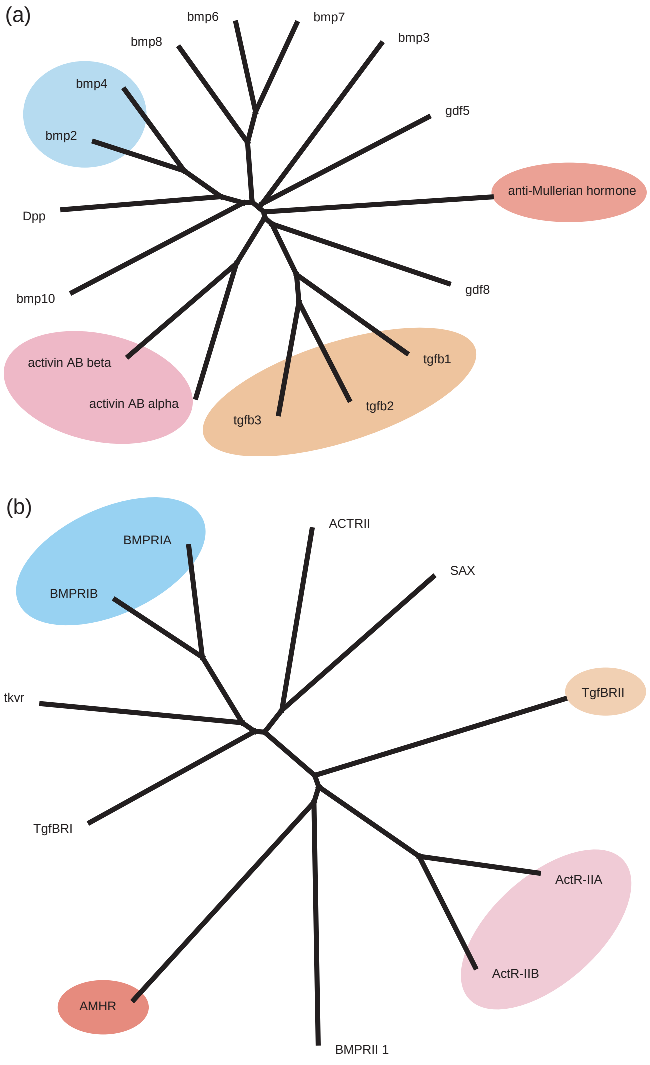

Fig. 1. Phylogenetic trees of the tgfβ ligands **(a)** and their receptors **(b)**. We have used matching colors for ligand–receptor pairs that are experimentally known to bind. The similar topologies of the trees demonstrate that the relative position of colored proteins are similar in both trees [e.g. the tgfβ ligands and activin ligands are nearby in (a) as are the respective tgfβII receptors and activinII receptors in (b)].

matches between two families allows us to identify potential interactions. The evolutionary relationships between the members of a protein family are reflected by a distance matrix. The distance matrix has dimensions $N \times N$, where N is the number of proteins in a family, and the (i, j) entry contains the evolutionary distance between protein i and protein j.

We have developed an algorithm to find the 'best fit' between two distance matrices by rearranging the rows and columns of one matrix in order to maximize the similarity between the two distance matrices. We initially assume a one-to-one correspondence between proteins in the first data set and proteins in the second data set, an assumption that is later removed. Because a protein family may contain 100 or more proteins, there are potentially $100! \approx 10^{158}$ permutations to examine. Since it is not feasible to conduct an exhaustive search, we employ an efficient approximate solution to solve the problem.

In this paper we report the application of our algorithm to match two families of ligands with their respective receptors: chemokines and tgfβ ligands. Chemokines are proteins that control diverse biological processes by activating G-protein-coupled receptors on the cell surface. These processes include angiogenesis, hematopoesis and organogenesis among many others. Understanding the interactions between chemokines and their receptors is an important first step in understanding the biology of these processes.

The tgfβ family of ligands and receptors contains a wide variety of proteins that are important in cellular differentiation and proliferation processes. Well-characterized members of the family include tgfβ, activins and the anti-Mullerian hormone. Mutations within these ligands have been implicated in a broad range of diseases such as cancer, diabetes, muscular dystrophy and hypertension. Bone morphogenic proteins (BMP) are other members of this family that induce bone formation by, among other pathways, stimulating osteocalcin in osteoblats.

2 METHODS

2.1 Experimentally derived interaction data sets

Our algorithm is applied to sets of chemokine ligands and receptors and tgfβ ligands and receptors. Experimentally determined interaction data between chemokines and their receptors in Table 1 is taken from the Database of Ligand–Receptor Pairs (DLRP) (Graeber and Eisenberg, 2001). Experimentally determined interaction data between tgfβ ligands and their receptors is contained in Table 4 (Massague and Chen, 2000; Massague, 1998).

2.2 Calculation of distance matrices

We obtained the proteins associated with the human chemokines and tgfβ ligands, along with their associated receptor families from Genbank. Using ClustalW (Higgins *et al.*, 1994) we first aligned the four protein families and then calculated the resulting distance matrices. The units of the distance matrix correspond to the percent difference in amino acid sequence.

2.3 Methodology

Given two distance matrices, we desire to find a 'best fit' between the two corresponding sets of protein sequences. A Monte Carlo method is used to select moves that maximize the similarity of one matrix with the other. After maximizing the fit we obtain the 'best' couplings between the two sets of proteins. Below we describe the algorithm in detail.

Downloaded from https://academic.oup.com/bioinformatics/article/19/16/2039/242166 by U.S. Department of Justice user on 17 August 2022

Table 1. Chemokine ligands and their receptors (DLRP) (Graeber and Eisenberg, 2001)

Chemokines	Receptors
GRO1	IL8RB
GRO2	IL8RB
GRO3	IL8RB
IL8	IL8RA, IL8RB
MIG	GPR9
PPBP	IL8RB
SCYA1	CCR8
SCYA2	CCBP2, CCR1, CCR2
SCYA3	CCR1, CCR5
SCYA4	CCR1, CCR5, CCR8, GPR9
SCYA5	CCBP2, CCR1, CCR3, CCR4, CCR5
SCYA7	CCBP2, CCR1, CCR2, CCR3, CCR5
SCYA8	CCBP2, CCR1, CCR2, CCR3, CCR5
SCYA11	CCBP2, CCR3, CCR5, GPR9
SCYA13	CCBP2, CCR1, CCR2, CCR3, CCR5
SCYA14	CCBP2, CCR1
SCYA15	CCR1, CCR3
SCYA17	CCR4, CCR8
SCYA19	CCR7
SCYA20	CCR6
SCYA21	CCR7, GPR9
SCYA22	CCR4
SCYA23	CCR1
SCYA24	CCR3
SCYA25	CCBP2
SCYA26	CCR3
SCYA27	
SCYB5	IL8RA, IL8RB
SCYB6	IL8RA, IL8RB
SCYB10	GPR9
SCYB11	GPR9
SCYB13	BLR1

Let X and Y be two distance matrices. The measure we consider is the correlation coefficient $r \in [-1, 1]$ given by

$$r(X,Y) := \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X_{ij}-\overline{X})(Y_{ij}-\overline{Y})}{\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(X_{ij}-\overline{X})^2}\sqrt{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(Y_{ij}-\overline{Y})^2}}$$

where $\overline{X}$ is the mean of all X_{ij}-values and $\overline{Y}$ is the mean of all Y_{ij}-values. When using the correlation coefficient, we are looking to maximize r, a unit-less measure. Our algorithm compares the upper-left $N \times N$ sub-matrix of X to all of Y, if X is larger than Y. When computing $r(X, Y)$, we only look at the entries of the larger matrix that are paired with proteins in the smaller set, these being the first N proteins in X.

The approach described so far assumes that each member of one family matches a single member of the interacting family. However, as we see in Tables 1 and 4 this assumption does not hold for the known ligand–receptor interactions, where multiple ligands bind multiple receptors. Therefore to more accurately capture the known ligand–receptor pairings, we find clusters of proteins that match between each data set and remove the one-to-one assumption.

To cluster proteins within a distance matrix we use the following technique known as Unweighted Pair Group Method with Arithmetic mean (UPGMA): (I) locate the smallest entry (i, j) in the matrix X (in the case of a tie, choose arbitrarily), (II) combine rows i and j and columns i and j by averaging their values (replace $X_{i\cdot}$ and $X_{\cdot i}$ with $(X_{i\cdot} + X_{\cdot i})/2$) and (III) delete row j and column j. At the end of this process the distance matrix ends up one dimension smaller.

The ‘cluster step’ just described can be used for two purposes. First, we can compare two different sized matrices by shrinking the larger one by repeatedly applying the cluster step until the data sets have the same dimension. Recall that under the one-to-one assumption, we were comparing a subset of the larger matrix with the entire smaller matrix which leads to a forced omission of data (due to a number of unused rows and columns). However, using the cluster steps allows the program to create one-to-many correspondences that include all of the proteins in the two families.

Moreover, we can also cluster both data sets to account for many-to-many relationships between the two families. This is important since as we see in Table 1, chemokines and chemokine receptors have many-to-many relationships. When performing this clustering, we compute a predetermined ‘cutoff’ for each matrix to decide when to stop the process of clustering. We take the cutoff for a matrix to be one standard deviation below the mean of the non-zero elements in the matrix. We simultaneously perform cluster steps on both matrices until the lowest entry in either matrix is above the cutoff for that particular matrix.

Once the clustering is complete, a Monte Carlo method is used to find the best fit between the two preprocessed matrices. The approach consists of an iterative improvement in the fitness function by random moves in the search space. A move is a randomly chosen change in the configuration of the system (Holm and Sanders, 1993). In our method, a move consists of choosing two indices i and j uniformly at random and then simultaneously swapping rows i, j and columns i, j. To decide whether such a move σ is accepted or rejected, we employ the Metropolis criterion (Metropolis *et al.*, 1953). The conditional probability $P(\sigma S_i|S_i)$ that a move from S_i to σS_i is accepted as the new configuration is

$$P(\sigma S_i|S_i) := \begin{cases} 1 & \text{if } \Delta E \le 0, \\ e^{-\Delta E/T} & \text{if } \Delta E > 0, \end{cases}$$

where ΔE (the ‘change in energy’) is the change in chosen measure of goodness of fit across the move and T is the ‘temperature’ of the system. With r as defined above, the change in energy is

$$\Delta E := r(X_i, Y) - r(\sigma X_i, Y).$$

Downloaded from https://academic.oup.com/bioinformatics/article/19/16/2039/242166 by U.S. Department of Justice user on 17 August 2022

Temperature T is the only parameter that needs to be specified when using this method. In order to help the random moves to settle into a minimum, the temperature is decreased according to an annealing schedule. To find an effective and efficient temperature function, we studied the heat capacity C of different systems representative of our problem. The heat capacity of a system is

$$C := \frac{\mathrm{d}\overline{E}}{\mathrm{d}T} = \frac{1}{T^2}\left(\overline{E^2} - \overline{E}^2\right).$$

To calculate C, we set the temperature constant at different levels and surveyed the energies of accepted moves. We hoped to observe a sharp increase in heat capacity when the temperature reached a critical value where the procedure could first be trapped in a local minimum. Such a sharp increase in C would be the result of the process having a tendency to fall into different minima and being trapped causing a large increase in the variation of energies. We did this study on systems varying in matrix size and degree of fit; specifically, we tested perfectly correlated matrices and completely random matrices. Each test using r consistently had a peak at about the same temperature, approximately 0.001.

Using the information from the heat capacity survey, we constructed a piecewise linear temperature function that makes many moves in the critical temperature range in order to minimize the likelihood of being trapped in a local minimum.

When the desired number of iterations (accepted or rejected) of the Monte Carlo procedure have been completed so that the optimal pairings S between the two data sets are in hand, the program calculates the overall 'goodness' of the pairings S by comparing it to a sampling of random sets of pairings so as to approximate a *z-score* for S. In order to find z, a vector of energies E_{rand} for random pairings between the data sets is first computed. The z-score is then

$$z := \frac{\overline{E_{\text{rand}}} - E_S}{\sigma_{E_{\text{rand}}}}$$

where $\sigma_{E_{\text{rand}}}$ is the standard deviation of $E_{\text{rand}}, \overline{E_{\text{rand}}}$ is the mean of E_{rand}, and E_S is the energy of S. A p-value can then be computed from

$$p = \frac{\operatorname{erfc}(|z/\sqrt{2}|)}{\sqrt{2}}$$

where erfc is the standard complement error function. This p-value is interpreted as the probability of randomly finding a pairing that has an equal or better fit than the pairing S found by the Monte Carlo optimization (Goh *et al.*, 2000). We report the values of the z-score and p for various distance matrices in Tables 2, 3 and 5.

We also test the goodness of individual pairings. To find this value, we calculate the correlation coefficient between the two row vectors that correspond to the two halves of the pairing. Pairings are output in decreasing order of individual correlation to suggest to the user which pairings are more likely to be true positive matches.

Table 2. Program-discovered one-to-one matches between chemokine receptors and ligands (50k iterations)

Receptors	$r = 0.983$	
	Ligands	Match correlation
IL8RA	SCYB5[a]	0.981
IL8RB	SCYB6[a]	0.975
CCR2	SCYA8[a]	0.967
CCR3	SCYA15[a]	0.967
CCR5	SCYA13[a]	0.946
CCR1	SCYA3[a]	0.892
CCR4	SCYA5[a]	0.886
CCR8	SCYA22	0.865
BLR1, CXCR5	SCYB10	0.639
CCR6	IL8	0.506
CXCR3, GPR9	SDF1	0.474
CCR7	SCYB13	0.375
CSCR4	SCYB11	0.318
CCBP2, CCR9	SCYA20	0.150
Z-score[b]	4.322	
p-value	1.0933×10^{-5}	

[a] Correct pairings from Table 1.

[b] The Z-score of obtaining a correlation of 0.983 by chance.

3 RESULTS AND DISCUSSION

3.1 Results

In the Monte Carlo phase, our MatLab Release 12 (9-22-2000) program can perform 10 000 iterations on a pair of data sets with 100 proteins each in about 25 s on a 733 MHz Intel Pentium III PC with 192 MB RAM. For 100 000 iterations, about 230 s are needed.

In the chemokine set we are working with, there are 63 known existing matches. We first tested the data under the one-to-one assumption. In this case the program is limited to finding only 14 possible pairings, this being the number of proteins in the smaller matrix. The best matches found with 50 000 iterations are shown in Table 2. Our program arrives at $r = 0.983$ between the two aligned distance matrices, with seven correct matches and a z-score of 4.1 for observing this many matches by chance.

We next computed the chemokine–receptor pairings in these data sets with our clustering technique enabled. Results for r are shown in Table 3. The program achieves $r = 0.946$, with 13 correct matches and a z-score of 3.2 for observing this many matches by chance.

Since these data sets are reasonably small, we did an exhaustive search to determine whether or not our program was finding the optimal pairings. Using the matrices attained by clustering, the exhaustive search found the same optimal

Downloaded from https://academic.oup.com/bioinformatics/article/19/16/2039/242166 by U.S. Department of Justice user on 17 August 2022

Table 3. Program-discovered cluster matches between chemokine receptors and ligands (50k iterations)

Receptors	$r = 0.946$	
	Ligands	Match correlation
CCR3	SCYA24[a]	0.870
CCR1	SCYA11	0.869
	SCYA2[a]	
	SCYA8[a]	
	SCYA13[a]	
	SCYA7[a]	
	SCYA1	
CCR5	SCYA14	0.875
	SCYA15	
	SCYA23	
	SCYA3[a]	
	SCYA4[a]	
	SCYA5[a]	
IL8RA	SCYB6[a]	0.825
	SCYB5[a]	
	GRO1	
	GRO2	
	GRO3	
	PPBP	
CCR4	SCYA26	0.804
CCR2	SCYA21	0.765
	SCYA19	
CCR8	SCYA22	0.730
	SCYA17[a]	
CXCR3, GPR9	SDF1	0.715
IL8RB	IL8[a]	0.681
BLR1, CXCR5	MIG	0.661
	SCYB10	
	SCYB11	
CCBP2, CCR9	SCYA25[a]	0.625
CCR6	SCYB13	0.397
CXCR4	SCYA27	0.331
CCR7	SCYA20	0.327
Z-score[b]	3.813	
p-value	9.7079×10^{-5}	

[a] Correct pairings from Table 1.
[b] The Z-score of obtaining a correlation of 0.946 by chance.

Table 4. Tgfβ ligands and their receptors (Massague and Chen, 2000; Massague, 1998)

Receptors	Ligands
TGFBRII	TGFB-1, TGFB-2, TGFB-3
ACTRIIa	ActivinβA, ActivinβB
ACTRIIb	ActivinβA, ActivinβB
AMHR	Anti-Mullerian hormone
BMPRIa	BMP-2, BMP-4, BMP-7, GDF-5
BMPRIb	BMP-2, BMP-4, BMP-7, GDF-5
SAX	Dpp
TKVR	Dpp

Table 5. Program-discovered cluster matches between tgfβ receptors and ligands (50k iterations)

Receptors	$r = 0.902$	
	Ligands	Match correlation
TGFBRII	tgfb2[a]	0.9146
	tgfb3[a]	
	tgfb1[a]	
BMPRIa	gdf5	0.9133
AMHR	Anti-Mullerian Hormone[a]	0.9055
BMPRIB	bmp2[a]	0.8990
	bmp4[a]	
ACTRIIa	ActivinβA[a]	0.8974
ACTRIIb	ActivinβB[a]	0.8922
SAX	bmp3	0.8650
TKVR	bmp10	0.8620
ACTRII	Dpp	0.8564
TGFBRI	bmp7	0.4677
	bmp6	
BMPRII	gdf8	0.25671
Z-score[b]	4.2474	
p-value	1.5292×10^{-5}	

[a] Correct known pairings from Table 4.
[b] The Z-score of obtaining a correlation of 0.902 by chance.

pairings for r as were achieved by our program. However the exhaustive search requires about 40 million steps, while the Monte Carlo search is able to find the same solution in only 50 000 steps.

To graphically demonstrate the effect of the Metropolis Monte Carlo optimization, we plotted the distance matrices for the clustered chemokines and the chemokine receptors before and after our method is applied (Fig. 2). From the graph it is clear that the matrices dramatically increase in similarity after the Metropolis Monte Carlo optimization.

We also computed the tgfβ pairings using the clustering technique. As seen in Table 4, there were 18 known matches, however our program is limited to finding a maximum of 11 matches (the number of proteins in the larger data set of the correct pairings). First, we only included proteins with known pairings to explore the accuracy achieved on these data sets. We recovered 10 out of 11 possible matches with $r = 0.965$ and a z-score of 3.3 for obtaining this many correct pairings by chance. Next, we added proteins that had no known pairings in order to predict matches. The results are shown in Table 5. The program found an $r = 0.902$ between the matrices, with eight correct matches out of 11. The z-score for observing this many matches by chance is 4.3.

With the chemokine set, all of the runs under the one-to-one assumption arrive at their respective solutions in about 185 s. With the clustering method, the runs take about 188 s each. The performance is about 3×10^{-7} s/proteins^2 iter. We did additional tests on the efficiency of the program by examining the convergence of r with the clustering method from the tests between the chemokine ligands and receptors. In each test,

Downloaded from https://academic.oup.com/bioinformatics/article/19/16/2039/242166 by U.S. Department of Justice user on 17 August 2022

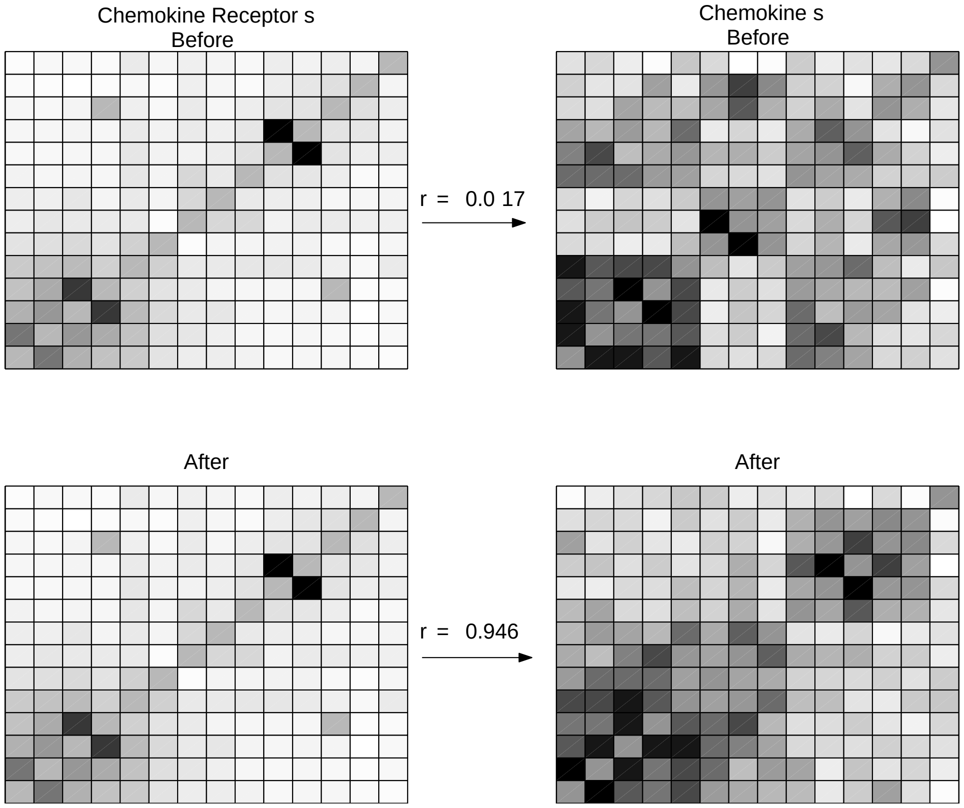

Fig. 2. A representation of the two distance matrices from the chemokine families with clustering before ($r = 0.017$) and after ($r = 0.946$) the Metropolis Monte Carlo optimization. The darker the box, the lower the corresponding value in the matrix.

we ran the program 100 times with the chemokine set for 50 000 iterations. We find that all runs converge to the same correlation.

As a control experiment, we also ran our program on the family of tgfβ ligands and chemokine receptors, to see whether we could correctly determine that there were no good matches between these families, as we would expect from experimentally derived interaction data. We found that the program was able to align these two distance matrices with a correlation of 0.84, which is lower than that achieved between the tgfβ ligands and receptors (0.97 with known pairings and 0.90 with additional family members) and the chemokines ligands and receptors (0.98 for one-to-one matches and 0.95 for clustered matches) distance matrices.

More significantly, we also found that the best pair-wise match between a tgfβ ligand and a chemokine receptor was 0.76. As seen in Tables 2, 3 and 5, this value is lower than the top matches found for both the chemokine–chemokine receptor pairings and the tgfβ–tgfβ receptor pairings. In fact, only 3 out of the 25 correct matches found in these tables have a correlation less than 0.76.

Therefore, comparing the results of aligning the tgfβ ligands to the chemokine receptors to those achieved between the tgfβ ligands and receptors and the chemokine ligands and receptors, we find that these two families generate lower correlation coefficients between the distance matrices and individual pairs of proteins. These results suggest that the methodology presented here can discriminate, in most of the cases examined here, between true and false positive interactions.

3.2 Discussion

Previous approaches have used the experimentally known matching of ligands and receptors to both evaluate the similarity of the two trees as well as to infer which uncharacterized ligands bind to which receptors (Goh and Cohen, 2002). To accomplish this one must first have a set of experimentally determined pairs of ligands and receptors. Then for each ligand, a set of distances is computed to the ligands in the experimentally determined pairs list. Similarly a set of distances is computed between each receptor and the receptors in the experimentally determined pairs list, maintaining the receptors and ligands in the same order from the pairs list. A correlation between the distances of any ligand and any receptor may then be computed. The receptors and ligand with higher correlation were shown to be more likely to interact than those with low correlation.

Downloaded from https://academic.oup.com/bioinformatics/article/19/16/2039/242166 by U.S. Department of Justice user on 17 August 2022

Although this approach is reasonable, it is more limited than our approach for various reasons. First, this approach requires prior knowledge of ligand receptor pairs, whereas our approach does not. Secondly, it treats each ligand and each receptor separately, whereas our methodology takes into consideration the matching between all clusters of ligands and all clusters of receptors simultaneously. However, due to our limited experimental knowledge of ligand–receptor pairs, it is difficult to quantitatively evaluate the different results produced by these two approaches.

The results described above demonstrate that using our simulated annealing approach we are able to recover about half of the possible ligand–receptor interactions for both the chemokine family of proteins and the tgfβ ligand–receptor pairs. The actual number of interactions recovered is greatly increased when we first cluster our protein families. This technique allows us to more closely model the one-to-many relationships found between the families that we explored.

It is important to note that our technique allows us to rank order predicted interactions by a match correlation. Therefore, although half the matches we predict are incorrect according to our current knowledge of interactions, the incorrect matches tend to fall among low scoring matches. Therefore, if we exclude the matches with low correlation, we in fact make predictions with accuracies significantly greater than 50%.

Finally, it should be noted that many of our supposedly incorrect matches may in fact be correct, but not yet experimentally demonstrated. One example of such a case occurs between the ligand SCYA11, or eotaxin, and receptor CCR1. Even though this interaction is not reported in the DLRP, we found that this ligand has been shown to bind the receptor CCR1 (Gao *et al*., 1996), and therefore this match is correctly predicted by our approach. This finding suggests that several other of our chemokine–receptor or tgfβ pairs may in fact represent true biological couplings that have not yet been discovered.

In conclusion, we have demonstrated that we can computationally discover ligand–receptor pairs. As such, this technique represents an addition to existing techniques that use computational methods to discover protein–protein interactions. In the future, we hope that these techniques motivate the experimental validation of the many protein interaction predictions that we are able to generate.

ACKNOWLEDGEMENTS

This project was funded by the National Science Foundation. The authors extend special thanks to the staff of IPAM and Protein Pathways and to Katherine Varley and Ron Elber.

REFERENCES

Gao,J.L., Sen,A.I., Kitaura,M., Yoshie,O., Rothenberg,M.E., Murphy,P.M. and Luster,A.D. (1996). Identification of a mouse eosinophil receptor for the CC chemokine eotaxin. *Biochem. Biophys. Res. Commun.*, **223**, 679.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Goh,C. and Cohen,F. (2002) Co-evolutionary analysis reveals insights into protein–protein interactions. *J. Mol. Biol.*, **324**, 177–192.

Goh,C., Bogan,A., Joachimiak,M., Walther,D. and Cohen,F. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.

Graeber,T.G. and Eisenberg,D. (2001) Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet.*, **29**, 295–300.

Higgins,D., Thompson,J., Gibson,T., Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Holm,L. and Sanders,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Massague,J. and Chen,Y.G. (2000) Controlling TGFbeta signaling. *Genes Dev.*, **14**, 627–644.

Massague,J. (1998) TGF-beta signal transduction. *Annu. Rev. Biochem.*, **67**, 753–791.

Metropolis,N., Rosenbluth,M.N., Rosenbluth,A., Teller,H. and Teller,E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087.

Overbeek,R., Fonstein,M., D’Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–28901.

Pazos,F. and Valencia,A. (2002) *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219–227.

Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Wagner,A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.*, **18**, 1283–1292.

Downloaded from https://academic.oup.com/bioinformatics/article/19/16/2039/242166 by U.S. Department of Justice user on 17 August 2022