doi:10.1006/jmbi.2000.3732 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 299, 283–293

JMB

COMMUNICATION

Co-evolution of Proteins with their Interaction Partners

Chern-Sing Goh[1], Andrew A. Bogan[2], Marcin Joachimiak[2] Dirk Walther[3] and Fred E. Cohen[3,4,5]*

[1]*Program in Medical Information Sciences*

[2]*Graduate Group in Biophysics*

[3]*Department of Cellular and Molecular Pharmacology*

[4]*Departments of Biochemistry and Biophysics and*

[5]*Medicine, University of California, San Francisco CA 94143, USA*

*Corresponding author

The divergent evolution of proteins in cellular signaling pathways requires ligands and their receptors to co-evolve, creating new pathways when a new receptor is activated by a new ligand. However, information about the evolution of binding specificity in ligand-receptor systems is difficult to glean from sequences alone. We have used phosphoglycerate kinase (PGK), an enzyme that forms its active site between its two domains, to develop a standard for measuring the co-evolution of interacting proteins. The N-terminal and C-terminal domains of PGK form the active site at their interface and are covalently linked. Therefore, they must have co-evolved to preserve enzyme function. By building two phylogenetic trees from multiple sequence alignments of each of the two domains of PGK, we have calculated a correlation coefficient for the two trees that quantifies the co-evolution of the two domains. The correlation coefficient for the trees of the two domains of PGK is 0.79, which establishes an upper bound for the co-evolution of a protein domain with its binding partner. The analysis is extended to ligands and their receptors, using the chemokines as a model. We show that the correlation between the chemokine ligand and receptor trees' distances is 0.57. The chemokine family of protein ligands and their G-protein coupled receptors have co-evolved so that each subgroup of chemokine ligands has a matching subgroup of chemokine receptors. The matching subfamilies of ligands and their receptors create a framework within which the ligands of orphan chemokine receptors can be more easily determined. This approach can be applied to a variety of ligand and receptor systems.

Keywords: co-evolution; protein interaction; ligand binding; G-protein coupled receptors; chemokines

The functions of proteins in biological systems are determined by the physical interactions they have with other molecules. Protein-protein binding is a subset of these interactions which is of primary importance in metabolic and signaling pathways. Proteins and their interaction partners must co-evolve so that any divergent changes in one partner's binding surface are complemented at the interface by their interaction partner (Atwell *et al.*, 1997; Jespers *et al.*, 1999; Moyle *et al.*, 1994; Pazos *et al.*, 1997). Otherwise, the interaction between the proteins is lost, along with its function. However, the co-evolution of interaction partners at the level of the whole protein family is not well understood. Most of our understanding of these interactions comes from genetic and biochemical experiments such as the common yeast two-hybrid assay (Fields & Song, 1989). Here, we consider if evolutionary information, in the form of statistical comparisons between the phylogenetic trees of protein families that interact with one another, can be used to recognize these interactions.

Recent advancements in using sequence information from completed genomes have improved the ability to predict general groups of interaction partners in the absence of experimental data using computational techniques. Two of these methods rely on gene fusion events to predict likely interacting genes, based on the assumption that genes that become fused into a single gene in any organism

C-S.G. and A.A.B. contributed equally to this work.

Present address: D. Walther, Incyte Genomics, Palo Alto, CA 94304, USA.

Abbreviations used: GPCR, G-protein coupled receptor; PGK, phosphoglycerate kinase.

E-mail address of the corresponding author: cohen@cmpharm.ucsf.edu

0022-2836/00/020283–11 $35.00/0

are likely to interact in other organisms (Enright *et al.*, 1999; Marcotte *et al.*, 1999a). Another approach has been to compare the presence and absence of homologous genes across multiple genomes to infer the involvement of a particular gene in a pathway involving other genes with similar profiles across multiple genomes (Pellegrini *et al.*, 1999). A combined algorithm that incorporates these approaches, and also messenger RNA expression comparisons, has recently been published (Marcotte *et al.*, 1999b). These approaches are quite useful for broadly defining functions of uncharacterized genes in completed genomes and for building general pathway information. However, they are not optimized to analyze the correlated divergent evolution of proteins and their interaction partners within a single ligand-receptor signaling system.

Ligand receptor systems often have multiple ligands that interact with a single receptor, or conversely, many receptors for a single ligand. To understand the co-evolution of a ligand gene family with its corresponding receptor gene family, it is necessary to quantify the correlated divergent evolution of the two families while including the biologically relevant pairings between ligands and receptors that are known to interact functionally. We have developed a method to measure quantitatively the correlation between the phylogenetic tree of a ligand family with the phylogenetic tree of a receptor family. The co-evolution of two interacting protein domains fused into a single gene was used to establish a guideline for analyzing the co-evolution of proteins and their interaction partners.

Co-evolution of domains in a single protein

The co-evolution of domains within a single protein is better understood than the co-evolution of proteins that are produced from different genes. Since domains within a single protein are covalently linked to one another by the polypeptide chain, the relationship between any two domains that interact with one another is one to one. We have chosen phosphoglycerate kinase (PGK) as a model system for quantifying co-evolution.

PGK is a two-domain protein with the enzyme active site formed by the interface between the two domains (Figure 1) (Banks *et al.*, 1979; Blake & Evans, 1974). PGK catalyzes the transfer of a phosphoryl-group from 1,3-bis-phosphoglycerate to ADP to form 3-phosphoglycerate and ATP, a critical step in glycolysis. A functional active site is achieved by the closing of the hinge between the two domains which positions the two substrates for the reaction (Bernstein *et al.*, 1997). Since the function of this enzyme depends on an active site formed between two independent domains, a working enzyme requires the two domains to have co-evolved. Any change in the N-terminal domain that perturbs the activity of the enzyme must be selected against, or subsequently compensated for, by a correlated change in the C-terminal domain. Because these two interacting domains are covalently linked, there is no ambiguity about each domain's interaction partner. For these reasons, PGK can be viewed as an example of co-evolution between two interacting domains. It is an ideal example for our statistical method of quantifying co-evolution between binding partners.

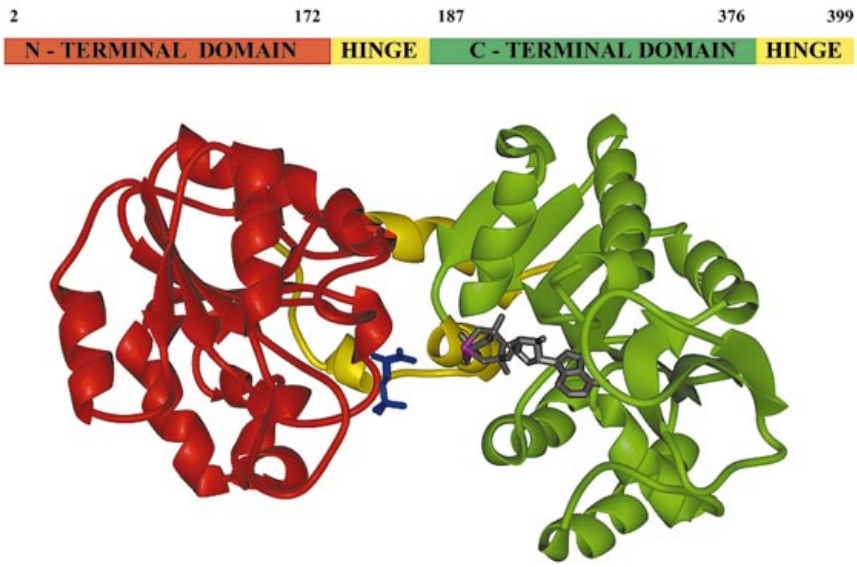

Figure 1. A ribbon diagram of the *T. maritima* PGK structure (PDB 1vpe). The N-terminal domain (residues 2-172) is in red, the C-terminal domain (residues 187-376) is in green, and the hinge regions (residues 173-186 and 377-399) are in yellow. The active PGK complex exhibits a hinge motion between the two terminal domains, bringing the two substrate ligands, 3-phosphoglycerate (blue) and ADP (gray) into close proximity (Bernstein *et al.*, 1997). The functional active site is formed at the interface of the two domains.

A multiple sequence alignment of PGKs from a vast array of species built with PSI-BLAST (Altschul *et al*., 1997) was divided into two independent alignments, one for the N-terminal domain and another for the C-terminal domain (Figure 1). The short linking regions, which are not directly involved in forming the active site, were left out of the two domain alignments. As a result, two phylogenetic trees were generated based on the pairwise sequence distances in the alignments, one tree for each domain (Figure 2). To quantify the similarity of the two trees we calculated the linear correlation coefficient between the set of all pairwise distances in tree 1 (N-terminal domain) with the equivalent distances in tree 2 (C-terminal domain) based on the actual covalent linkages between the domains (see Methods). For the N and C-terminal domain trees, the correlation coefficient was 0.79(±0.01), with a *z*-score of 41.91 (Table 1), indicating that the divergent evolution of the N termini from one another is highly correlated with the divergent evolution of the C termini from one another.

To validate that this correlation was a meaningful measure of the co-evolution of the two domains, we recalculated the correlation coefficient using randomly chosen incorrect pairings between the domains. N and C-terminal domains from a single PGK gene were therefore not paired with one another, but were incorrectly matched with a domain from a different PGK. The correlation coefficient between the trees for these non-binding pairs was 0.00(±0.02), with a *z*-score of 0.29 (Table 1). The lack of correlation between mismatched pairs serves as a control for our analysis method and shows that the correct linkage of domains with their real binding partners is required to observe co-evolution. To control further for the effects of speciation, as opposed to co-evolution, we also calculated the correlation coefficient between the tree for full-length PGKs from 17 different species and a tree for topoisomerases (an enzyme that does not interact with PGK) from the same 17 species. The correlation coefficient for these two trees is 0.54(±0.08) with a *z*-score of 6.25. This lower correlation coefficient suggests that, while speciation is an important effect, the higher correlation between the trees of the PGK N and C-terminal domains is due to co-evolution and not just speciation.

The quantitative recognition of the co-evolution of the two domains of PGK was fully expected, since the two domains are linked to one another and must interact in order to function as an enzyme. However, a perfect correlation was not seen, since irregularities in the coordinated evolution of a single gene do occur, albeit relatively infrequently. For example, gene duplication or acquisition followed by domain swapping might allow for pairings of N and C-terminal domains that did not diverge together. It appears that this type of unexpected pairing of distantly related domains has occurred in the black spruce tree *Picea mariana*. Its PGK C-terminal domain clusters with those of other closely related viridiplantae whose PGKs appear to derive from a eubacterial lineage (Figure 2(b)). However, the N-terminal domain of *P. mariana* PGK is more similar to the eukaryotic alveolata than to the other viridiplantae N termini, which remain with the eubacterial lineage (Figure 2(a)). The clustering of viridiplantae and euglenozoa PGKs within the eubacterial lineage (Figure 2, in green and pink) suggests that, in those groups of eukaryotes, PGK has most likely evolved from the genetic material of an organelle with eubacterial origins.

For a two-domain protein such as PGK, most of these domain swapping events are selected against, since function is rarely preserved. *P. mariana* PGK is clearly an exception, not the rule. A few other organisms, such as *Drosophila melanogaster* and *Plasmodium falciparum*, show poor correlation between the two PGK domains in Figure 2, but the vast majority have clearly co-evolved. We conclude that a reasonable upper bound for a correlation coefficient in a system that has co-evolved is approximately 0.8. With this standard in mind from the PGK example, it is possible to evaluate the co-evolution of more complicated systems, such as ligands and their receptors.

Table 1. Correlation coefficients and related statistics

A. *PGK N terminus and PGK C terminus*	
Binding pairs	
Correlation coefficient:	0.79 ± 0.01
z-score:	41.91
p-value:	0.00
Non-binding pairs	
Correlation coefficient:	0.00 ± 0.02
z-score:	0.29
p-value:	0.77
B. *Chemokines and chemokine receptors*	
Binding pairs	
Correlation coefficient:	0.57 ± 0.02
z-score:	21.82
p-value:	0.00
Non-binding pairs	
Correlation coefficient:	0.01 ± 0.03
z-score:	0.41
p-value:	0.68
C. *Human-only chemokines and chemokine receptors*	
Binding pairs:	
Correlation coefficient:	0.44 ± 0.04
z-score:	11.23
p-value:	2.87×10^{-29}
D. *PGKs and Topoisomerases*	
Species pairs:	
Correlation coefficient:	0.54 ± 0.08
z-score:	6.25
p-value:	3.92×10^{-10}

Binding pairs refer to the pairs of interacting partners used in our statistical analysis (see Methods). They are either covalently linked (in the case of PGKs two domains) or experimentally known to bind one another (in the case of the chemokines and their receptors). Non-binding pairs were chosen at random and are not believed to interact. Since PGKs and topoisomerases do not bind to one another, pairings were done by species.

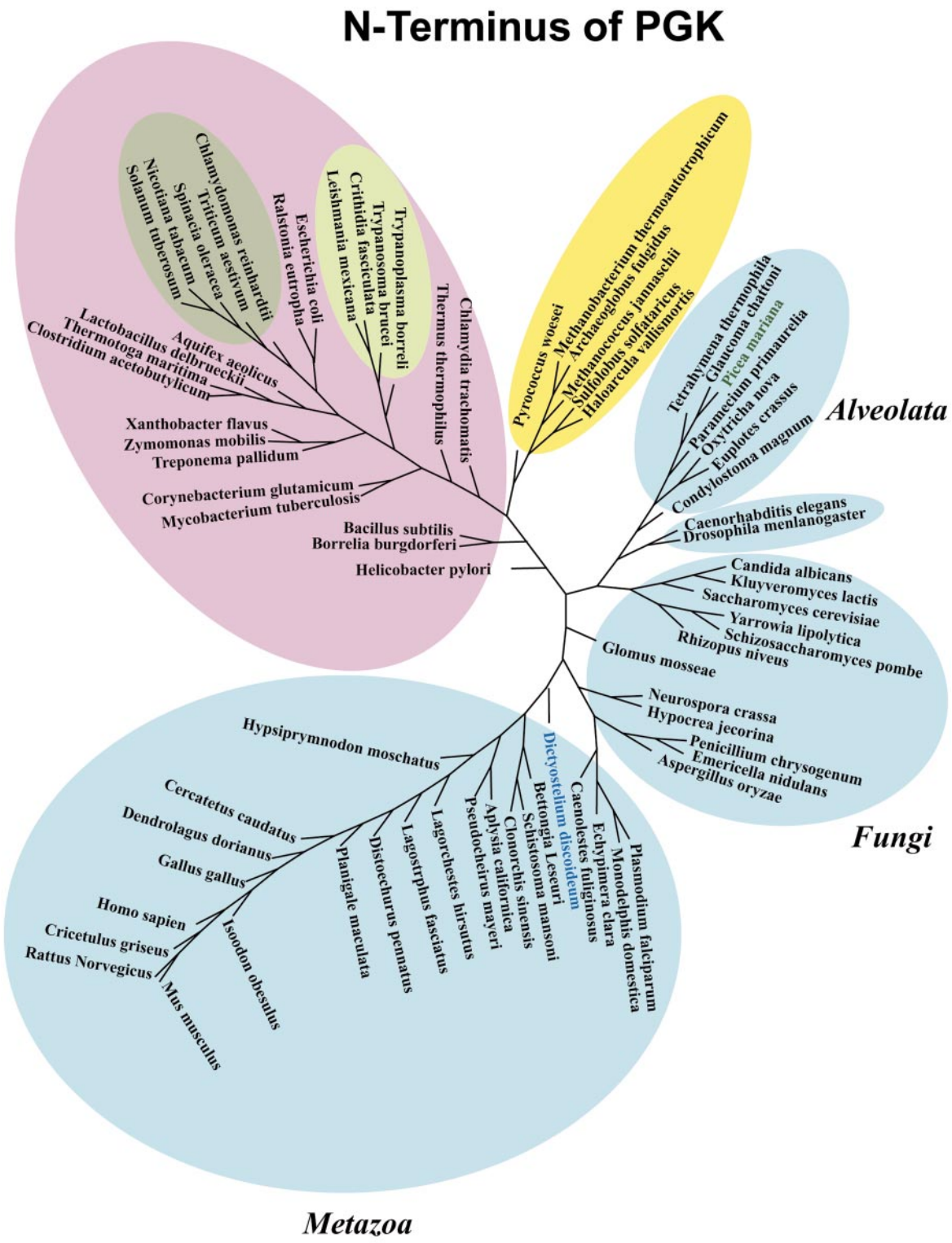

Figure 2 (*legend opposite*)

Co-evolution of ligands and receptors

Ligands and receptors, as with interacting domains, must co-evolve, both to preserve necessary signaling pathways and to allow for the creation of new pathways during the evolution of an organism. However, it has been quite difficult to quantify or visualize the co-evolution of ligands and their receptors. We have applied our technique for measuring co-evolution to a ligand-receptor

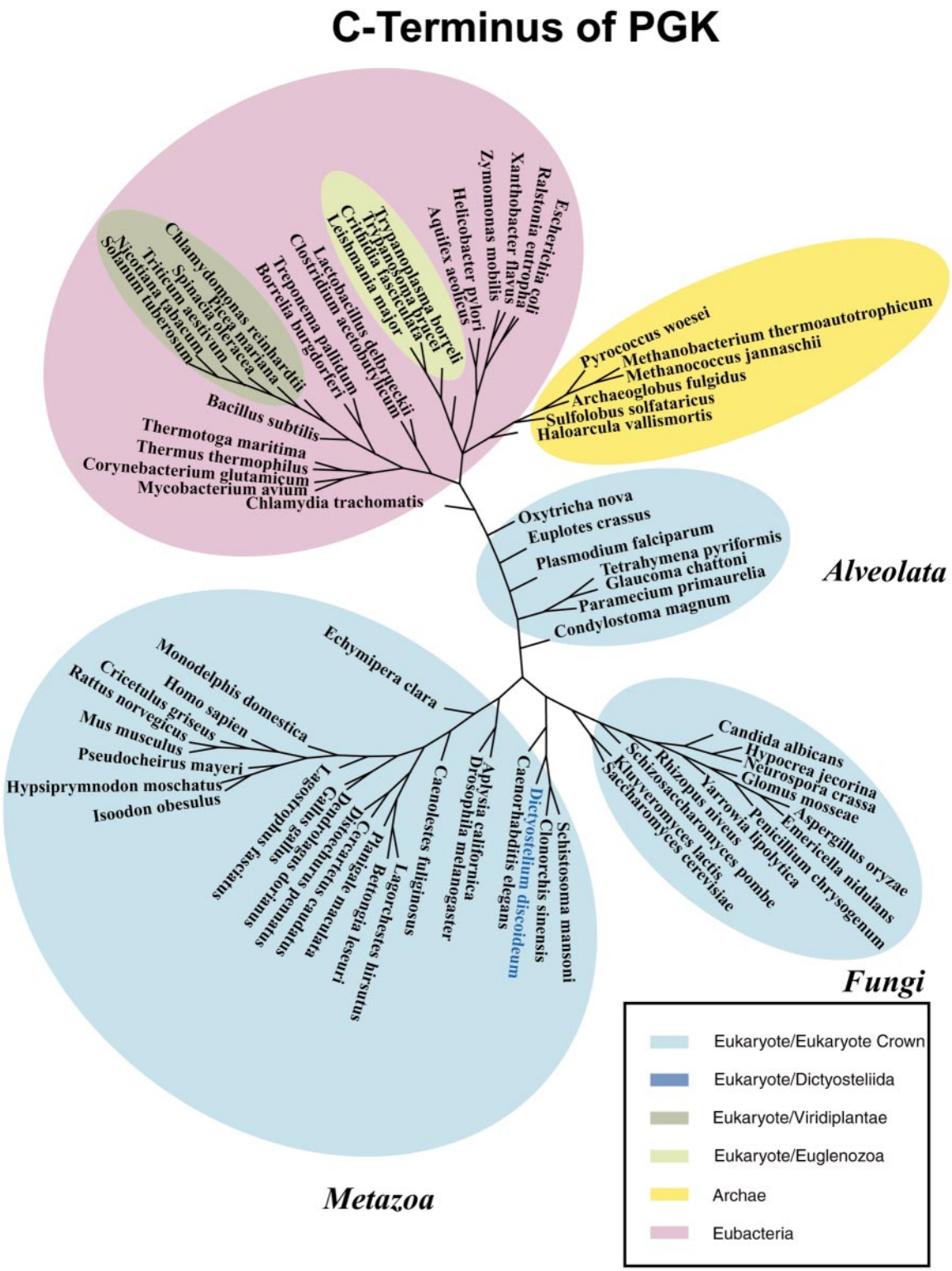

Figure 2. The phylogenetic trees of the N-terminal and C-terminal domains of PGK. (a) N-terminal domains and (b) C-terminal domains of PGK cluster into separate kingdoms of eukaryotes (blue), eubacteria (pink), and archae (yellow). The eukaryotic groups of viridiplantae and euglenozoa cluster among the eubacteria sequences indicating that, for this enzyme, these sequences are evolutionarily closer to orthologs in eubacteria than to orthologs in other eukaryotes.

system that is well suited for this analysis, the chemokines and their transmembrane receptors. This is good model system for relating primary sequence knowledge to biological function. Our goal was to obtain information relevant to ligand-receptor binding specificity from sequence data.

Chemokines constitute a large family of *chemo*tactic cyto*kines* that activate transmembrane G-protein-coupled receptors (GPCRs) on the cell surface to regulate diverse biological processes. These processes include leukocyte trafficking, angiogenesis, hematopoiesis, and organogenesis (Baggiolini *et al.*, 1997; Oppenheim *et al.*, 1991). Chemokines are believed to be both beneficial in host defense against infectious agents and harmful in diseases marked by pathologic inflammation. All nucleated cells are capable of expressing at least some chemokines, and it appears that these molecules perform an extracellular messenger role in all tissues and systems of the body (Locati & Murphy, 1999). The chemokines are found in higher vertebrates and the ones included in this study are from various mammals (human, monkey, rat, mouse, pig, guinea pig, cow, sheep, dog, horse, rabbit, mangabey, gorilla, and chimpanzee), frog, and chicken.

Recently, there has been increasing interest in chemokine receptors because CXCR4 and CCR5 have been found to be co-receptors for CD4-mediated HIV entry into cells (Premack & Schall, 1996). Not only do chemokines play a pivotal role in HIV infection, but they also exert other effects in inflammatory conditions and cancer (Wang *et al.*, 1998). Targeting specific chemokines and chemokine receptors may have therapeutic utility in inflammation, cancer, and infectious disease. The important role of chemokine signaling in disease, coupled with the wide variety of known chemokines and chemokine receptors, render this system ideal for studying the co-evolution of ligands and their receptors.

The chemokine nomenclature is defined by a cysteine residue signature motif where C is a cysteine residue and X is any amino acid residue (Clore & Gronenborn, 1995). They fall into four categories: CXC, CC, C, and CX_3C. Most of the known chemokines are members of the CXC or CC subfamilies. The C and the CX_3C chemokine subfamilies were discovered more recently. The first C chemokine found was lymphotactin; fractalkine was the first CX_3C chemokine discovered (Bazan *et al.*, 1997; Kelner *et al.*, 1994). We have selected various chemokine receptors and their cognate ligands for this analysis (Table 2).

Our technique for mapping and quantifying the co-evolution of binding specificity was applied to the chemokine system. We built trees that show the correlated evolution of binding specificity for chemokines and their receptors (Figure 3). Using the known information regarding the binding of chemokines and their cognate receptors (Table 2) we calculated the correlation coefficient for the chemokine ligand and receptor trees. The correlation coefficient for these trees is 0.57(±0.02) with a *z*-score of 21.82 (Table 1). Considering the upper bound of 0.8, which we have established using PGK, a two-domain system that has clearly co-evolved, the correlation coefficient of 0.57 indicates a very highly correlated co-evolution of the chemokines and their receptors. Since very few different (and less divergent) species were used in this case, the effects of speciation are much less significant for the chemokine system than they were for the PGK example. Still, we confirmed that speciation was not a major factor by calculating the correlation coefficient between the chemokines and their receptors within a single species. For only the human chemokines and their receptors, the correlation coefficient between the trees is 0.44(±0.04) with a *z*-score of 11.23 and a *p*-value of 2.87×10^{-29}.

For any given chemokine, its closest sequence neighbors are far more likely to bind the closest neighbors of its receptor than to bind a randomly selected chemokine receptor. The analysis applies to all the chemokines in the phylogenetic tree (Figure 3) based on their known binding partners (Table 2). Our all-inclusive approach and calculation of a statistical correlation coefficient may explain why we find a high degree of co-evolution despite a previous study that concluded that CC chemokines had not co-evolved closely with their receptors (Hughes & Yeager, 1999). Our control calculation was done based on incorrect binding partners chosen at random. For this random, non-binding map of ligands to receptors, the correlation coefficient was 0.01(±0.03), with a *z*-score of 0.41 (Table 1). The non-correlation of randomly paired ligands and receptors demonstrates that the real biological interaction partners must be chosen to show co-evolution between ligands and their receptors. Since it is easy to add new sequences to phylogenetic trees, our approach creates a scalable

Table 2. Chemokine receptors and their ligands

CC chemokine receptors	CC chemokines
CCR1	MIP1α RANTES, MCP3, HCC1, MPIF1, MIP5
CCR2	MCP1, MCP2, MCP3, MCP4, MCP5
CCR3	Eotaxin, MCP2, MCP3, MCP4, RANTES, Eotaxin2, MIP5
CCR4	TARC, MDC
CCR5	MIP1α, MIP1β, RANTES
CCR6	MIP3α
CCR7	MIP3β, SLC
CCR8	I-309, TARC, MIP1β
CCR9	TECK
CXC chemokine receptors	CXC chemokines
CXCR1	IL-8
CXCR2	IL-8, GCP2, GRO-α, β, γ, ENA78, PGP
CXCR3	IP10, MIG
CXCR4	SDF1
CXCR5	BLC
C chemokine receptor	C chemokine
XCR1	Lymphotactin
CX_3C chemokine receptor	CX_3C chemokine
CX3CR1	Fractalkine

These experimentally determined binding partners (Baggiolini *et al.*, 1997; Kim & Broxmeyer, 1999; Lu *et al.*, 1999; Rollins, 1997; Zaballos *et al.*, 1999) were used to calculate the correlation coefficient between the ligand and receptor trees (see Methods).

framework allowing new chemokine or receptor sequences to be clustered based on their likely binding specificity. The search space for experimental determination of a novel family members' interaction partners is therefore greatly reduced. More detailed information about the binding specificity of the chemokines and their receptors can be obtained by analyzing the correlated phylogenetic trees (Figure 3).

Analysis of chemokine co-evolution

In Figure 3(b), the CXC receptors cluster in a separate group from the CC receptors, with the C and CX_3C receptors forming their own group roughly equidistant from the CXC clusters and the main two groups of CC receptors. Among the CC receptors, CCR1, CCR2, CCR3, and CCR5 have sequences that are closely related to one another. CCR4 and CCR8 cluster together, as do CCR6, CCR7, CCR9, and the orphan receptor STRL33. This last subset of CC receptors falls as close to the CXC receptors as it does to the C and CX_3C receptors. Correspondingly, the ligands of the chemokine receptors form clusters that match the branches of the receptor tree (Figure 3(a)).

It is important to note, that there is some subjectivity in the assignments of clusters on the two trees (Figure 3). We have attempted to choose groupings that correspond to known physiological interactions wherever possible. For example, since CCR4 and CCR8 share a common ligand, TARC, we have chosen to group CCR4 and CCR8 together instead of grouping CCR8 with CX_3CR1 (an equally plausible cluster-based on the tree alone). However, these arbitrary choices were not used in the calculations of the correlation coefficients and therefore do not impact our statistical data.

The MIP chemokines (except MIP3) and RANTES group together, as do the nearby MCP chemokines and eotaxin (Figure 3(a), colored pink). Subsets of these chemokines bind to CCR1, CCR2, CCR3, and CCR5 (Table 2), which form a cluster on the receptor tree (Figure 3(b), also in pink). Similarly, MIP3α, MIP3β, TECK, and SLC cluster together (Figure 3(a), in light red). MIP3α binds to CCR6; while MIP3β and SLC bind to CCR7. TECK binds to CCR9. The corresponding cluster can be found on the receptor tree where CCR6, CCR7, and CCR9 form a third subgroup of CC receptors along with the human orphan chemokine receptor STRL33 (Figure 3(b), in light red).

Within the CXC chemokine receptors, CXCR1 and CXCR2 group together (Figure 3(b), in green). CXCR1 binds to IL-8; and CXCR2, with its broader specificity binds to IL-8, GCP2, the GROs, ENA78, and PGP. On the ligand tree, these chemokines also form a cluster within the other CXC chemokines (Figure 3(a), in green). CXCR3, on its own branch of the CXC receptor cluster, binds to MIG and IP10, which cluster together on the chemokine tree (Figure 3, in blue). The human chemokine H174 also falls in this group. CXCR4 binds to SDF1 (Figure 3, in yellow) and CXCR5 binds to BLC (Figure 3, in magenta). The branching structure of the CXCR3-5 branches (Figure 3(b), in blue, magenta, and yellow) is not, however, identical with the branching structure of their ligands (Figure 3(a), in blue, magenta, and yellow). While the clusters still match between the trees, these differences in the branching patterns contribute to the imperfect correlation between the trees.

The grouping of the C and CX_3C chemokine receptors on the receptor tree also corresponds with their ligands. The C chemokine, lymphotactin, and the CX_3C chemokine, fractalkine, can be grouped on the chemokine tree (Figure 3, in gray). This implies that the binding specificities of these two types of receptor are closer to one another than to CC or CXC receptors. However, because there is only one example of each of these two classes of chemokine receptors, there may be some bias toward pairing these sequences. Therefore, the C and CX_3C chemokines and their receptors may be less closely related than they appear on the trees.

Since the chemokine and receptor trees cluster according to their binding specificities, we can begin to make inferences about possible ligands for orphan receptors and *vice versa* (the ''orphan'' designation means that a cognate ligand or a cognate receptor is not known for a receptor or chemokine, respectively). Several orphan chemokines and one orphan chemokine receptor were included in the trees (Figure 3). The orphan receptor STRL33 (Liao *et al*., 1997) groups with CCR6 and CCR7. Based on the high correlation coefficient for our trees, we suggest that the orphan receptor STRL33 is likely to bind a chemokine that is from the corresponding group on the chemokine tree. This suggests that likely ligand candidates are chemokines (either known or not yet discovered) related to MIP3α, MIP3β, SLC, or TECK.

The human chemokine H174, which at the start of this study was an orphan, clusters with MIG and IPl0 (Figure 3(a), in blue), so we suggested that H174 binds a CXC chemokine receptor, most likely CXCR3 or one that is very similar in sequence. A recent independent experimental study has confirmed this prediction showing that H174 (also known as IP-9) is a high-affinity ligand for CXCR3 (Tensen *et al*., 1999). Two other orphan chemokines, HCC4 and MIP4 (Guan *et al*., 1999; Hedrick *et al*., 1998), cluster with their related CC chemokines (Figure 3(a), in pink). We predict that the receptors of these orphan chemokines are likely to fall within the pink cluster of CCR receptors in Figure 3(b).

PF4, another orphan chemokine, clusters with the ligands of CXCR1 and CXCR2. However, it is known that PF4 does not bind CXCR1 or CXCR2 in its wild-type form. Interestingly, engineered protein constructs containing a modification of the

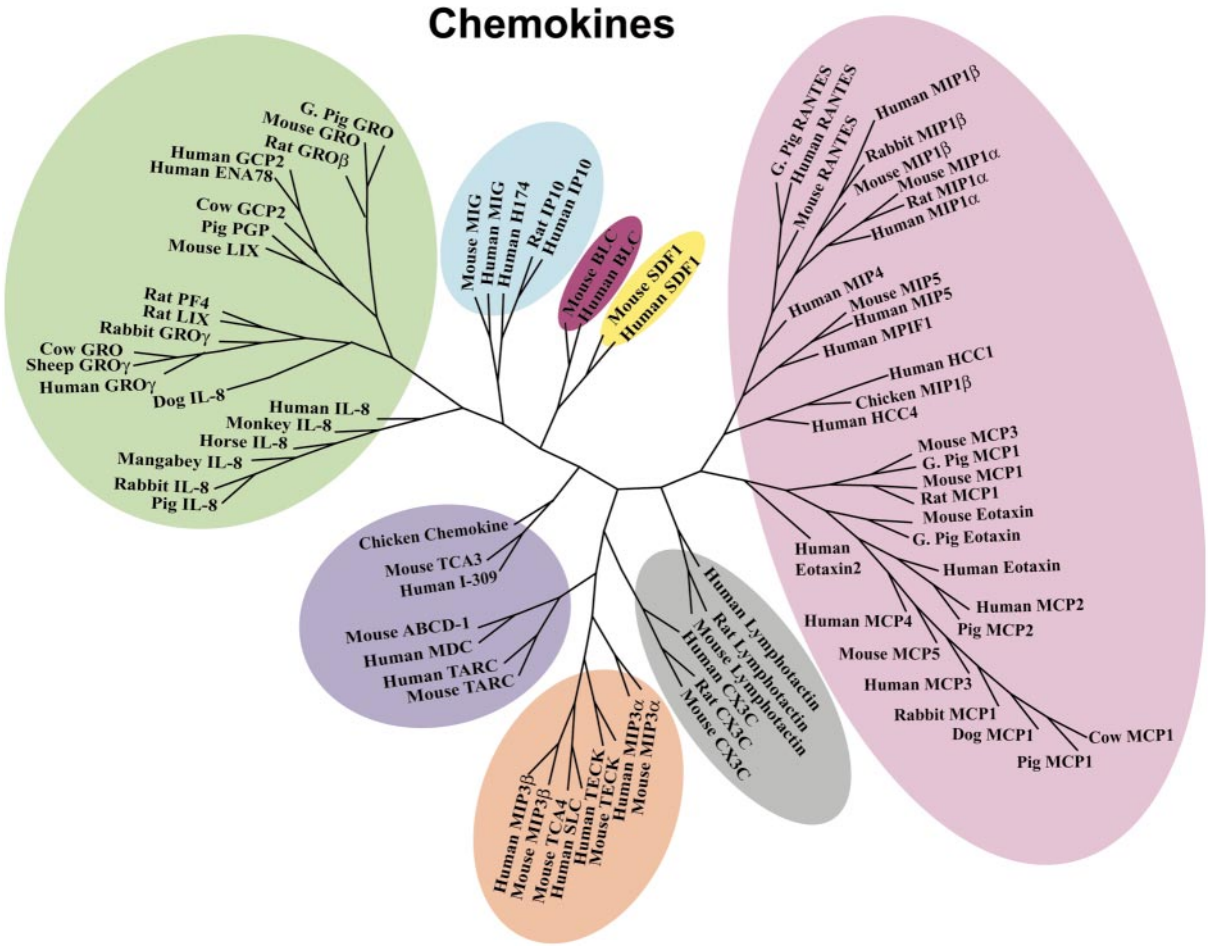

Figure 3 (*legend opposite*)

N-terminal sequence of PF4 do bind to CXCR2 (Jones *et al*., 1997). This implies that the sequence is competent for the predicted specificity, but its potential to interact has been suppressed by divergent evolution within specific regions of its N terminus. In the case of PF4, the oligomerization state of the chemokine may control its biological function. A recent study shows that tetrameric PF4 binds directly to glycosaminoglycans on the surface of neutrophils (Peters *et al*., 1999).

Conclusions

The co-evolution of the two domains of phosphoglycerate kinase was used to develop a guideline for quantifying co-evolution of proteins and their binding partners. Based on this guideline, the chemokines and their receptors were shown to have co-evolved. Our method was applied to orphan ligands and receptors in the search for orphans' binding partners. It provides a framework that significantly reduces the search space from all possible ligands or receptors to a small subset represented by a region of our phylogenetic tree. While the binding interactions of orphan ligands and receptors can only be proven experimentally, this analysis should aid in the rapid discovery of currently unknown chemokine signaling pathways.

The approach is readily expandable to include new ligand and receptor sequences as they are discovered. It can also be applied to other systems of proteins and their interaction partners. Possible examples include other cytokines and kinases. It is also potentially useful for representing the evolution of ligand binding specificity in systems that have small-molecule ligands, such as nuclear hormone receptors and other GPCRs once a suitable phylogeny of small molecules or the enzymes responsible for their biosynthesis can be established.

Methods

Sequence analysis

Sequences related to human CXCR1, IL-8, and phosphoglycerate kinase were retrieved using PSI-

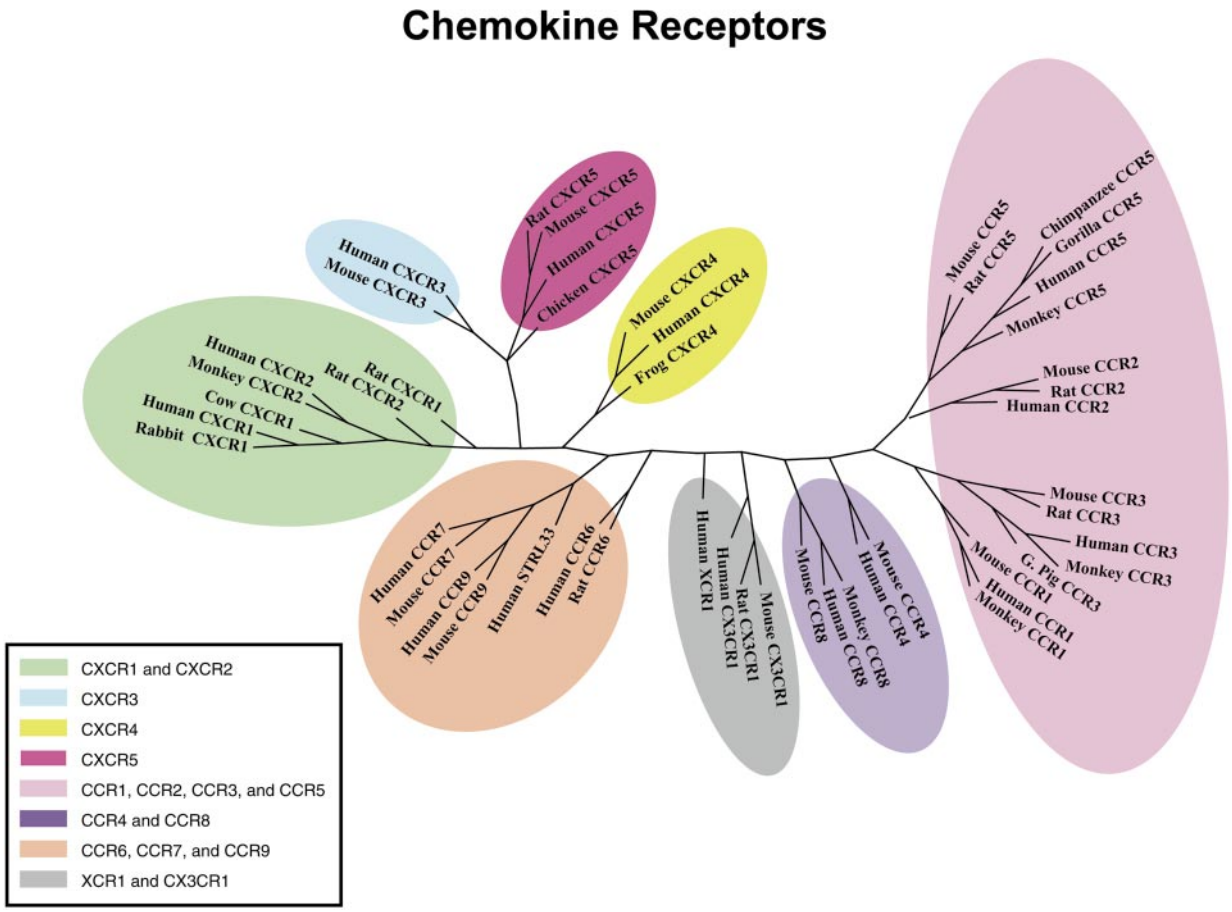

Figure 3. Phylogenetic trees of (a) chemokines and (b) chemokine receptors. The diagrams are colored by their clustering patterns to show similar groupings among the chemokines and the receptors to which they bind. The colored groups were chosen by eye based on the branching of the chemokine receptor tree. They are provided only as a guide for visualization of the data and were not used in the calculation of the correlation coefficients.

BLAST (Altschul *et al.*, 1997) with default parameters and the complete non-redundant database. Multiple sequence alignments of the chemokine receptors, the chemokines, and the phosphoglycerate kinases were constructed based directly on the PSI-BLAST alignments. The multiple sequence alignment for PGK was divided into two alignments, one for each domain. The N-terminal domain alignment included residues 2-172 and the C-terminal domain included residues 187-376. Topoisomerase I sequences from 17 different species (including eukaryotes, eubacteria, and archae) were selected from the SWISSPROT database and aligned using ClustalW. The ClustalW phylogeny program was used to calculate a distance matrix by percentage sequence divergence and to generate the trees with the neighbor-joining method (Saitou & Nei, 1987). The unrooted trees were drawn using the DrawTree program in PHYLIP (Felsenstein, 1993).

Correlation analysis

Distance matrices were generated from the multiple alignments using ClustalW (Thompson *et al.*, 1994). In order to quantify the co-evolution of interaction partners, we employed a linear regression analysis measuring the correlation between pairwise evolutionary distances among all proteins in a multiple sequence alignment. These were correlated with the evolutionary distances among the corresponding binding partners (or, in the case of PGK and topoisomerase I, the corresponding species, since these proteins do not bind). We defined X as a two-dimensional matrix of evolutionary distances in the receptor family (X was constructed as a $N \times N$ matrix, where N is equal to the number of receptors). For the corresponding ligands, a similar distance matrix, Y, was constructed. X_{ij} is the pairwise distance between sequence m_i and sequence m_j. Y_{ij} signifies the pairwise distance between sequence n_i and sequence n_j

(where n_i is experimentally known to bind to m_i and n_j is known to bind to m_j). In order to represent multiple ligands that bind to a single receptor, or *vice versa*, there were instances where the same ligand or receptor was represented more than once in the matrix. Therefore, in the cases where one ligand was known to bind experimentally to two different receptors, the ligand was represented as both n_i and n_j in matrix Y, corresponding to the two different receptors, m_i and m_j, in matrix X. The correlation coefficient was then calculated for all the pairwise distances in matrix X and their corresponding distances in matrix Y.

We computed the linear correlation coefficient r (Pearson's correlation coefficient (Press *et al.*, 1988)) defined as:

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

with $-1 \leqslant r \leqslant +1$ where $\bar{X}$ is the mean of all X_{ij}-values and $\bar{Y}$ is the mean of all Y_{ij}-values. In our context, X_{ij} and Y_{ij} are pairwise sequence similarity distances between N-terminal and C-terminal domains of PGK, or between chemokine receptors and their corresponding chemokines, respectively. Positive values of r would indicate a positive co-evolution; i.e. receptors that appear to be evolutionarily close, have ligands that, in turn, are more closely related than other pairs of any two ligands. By contrast, r-values of around zero would indicate no correlation, and negative values of r would indicate anti-correlation.

Estimation of statistical significance of correlation

The significance of the computed value r was assessed by a bootstrapping analysis yielding an estimate of the standard deviation of r given the size of our data set (Efron, 1979), and by an estimation of the probability of obtaining the observed value of r by chance (p-value). In the bootstrap analysis, we generated 1000 sets containing N pairwise distances randomly drawn (with replacement) from the N pairwise distances in the original set. For every such set we computed the bootstrap correlation coefficient r_b. The bootstrap interval, i.e. the interval of r_b accounting for 68 % of the obtained values of r_b was obtained from the 16 % (a) and 84 % (b) percentiles in the histogram of the 1000 values of r_b and the mean value of r_b from the 50 % percentile. The bootstrap estimate of the standard deviation of the observed correlation then calculates as:

$$\sigma_b = \frac{b-a}{2}$$

The p-value, i.e. the probability that the particular correlation coefficient r, quantifying the co-evolution between chemokines and their receptors was obtained by chance, was obtained by randomly shuffling the pairwise distances between ligands and receptors. Thus the assignments of correspondence (ligand l_1 binds to receptor R_{l1}, and ligand l_2 binds to receptor R_{l2}) were replaced by random assignments, and the correlation coefficient was computed as explained above. This process was repeated 1000 times. From the resulting 1000 values r_{rand}, a z-score for the actual observed value r was calculated as:

$$z = \frac{r - \bar{r}_{rand}}{\sigma_{rand}}$$

where σ is the standard deviation of r_{rand} and $\bar{r}_{rand}$ is the mean (effectively zero for truly random data). The p-value is then obtained from $p = erfc(|z|)/\sqrt{2}$, where $erfc$ is the complement error function.

Acknowledgments

A.A.B. was supported by a National Defense Science and Engineering Graduate Fellowship from the United States Department of Defense and by the Lloyd M. Kozloff Fellowship. This work was supported by grants from the NIH (to F.E.C.) and additional funding was provided by Pfizer. We thank Jonathan Blake and John-Marc Chandonia for helpful discussions.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Atwell, S., Ultsch, M., De Vos, A. M. & Wells, J. A. (1997). Structural plasticity in a remodeled protein-protein interface. *Science,* **278**, 1125-1128.

Baggiolini, M., Dewald, B. & Moser, B. (1997). Human chemokines: an update. *Annu. Rev. Immunol.* **15**, 675-705.

Banks, R. D., Blake, C. C., Evans, P. R., Haser, R., Rice, D. W., Hardy, G. W., Merrett, M. & Phillips, A. W. (1979). Sequence, structure and activity of phosphoglycerate kinase: a possible hinge-bending enzyme. *Nature,* **279**, 773-777.

Bazan, J. F., Bacon, K. B., Hardiman, G., Wang, W., Soo, K., Rossi, D., Greaves, D. R., Zlotnik, A. & Schall, T. J. (1997). A new class of membrane-bound chemokine with a CX3C motif. *Nature,* **385**, 640-644.

Bernstein, B. E., Michels, P. A. & Hol, W. G. (1997). Synergistic effects of substrate-induced conformational changes in phosphoglycerate kinase activation. *Nature,* **385**, 275-278.

Blake, C. C. & Evans, P. R. (1974). Structure of horse muscle phosphoglycerate kinase. Some results on the chain conformation, substrate binding and evolution of the molecule from a 3 angstrom Fourier map. *J. Mol. Biol.* **84**, 585-601.

Clore, G. M. & Gronenborn, A. M. (1995). Three-dimensional structures of alpha and beta chemokines. *FASEB J.* **9**, 57-62.

Efron, B. (1979). Computers and the theory of statistics: thinking the unthinkable. *Soc. Ind. Appl. Math. Rev.* **21**, 460-480.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature,* **402**, 86-90.

Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package)*, 3.5c edit., Department of Genetics, University of Washington, Seattle.

Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature,* **340**, 245-246.

Guan, P., Burghes, A. H., Cunningham, A., Lira, P., Brissette, W. H., Neote, K. & McColl, S. R. (1999). Genomic organization and biological characterization of the novel human CC chemokine DC-CK-1/PARC/MIP-4/SCYA18. *Genomics,* **56**, 296-302.

Hedrick, J. A., Helms, A., Vicari, A. & Zlotnik, A. (1998). Characterization of a novel CC chemokine, HCC-4, whose expression is increased by interleukin-10. *Blood,* **91**, 4242-4247.

Hughes, A. L. & Yeager, M. (1999). Coevolution of the mammalian chemokines and their receptors. *Immunogenetics,* **49**, 115-124.

Jespers, L., Lijnen, H. R., Vanwetswinkel, S., Van Hoef, B., Brepoels, K., Cohen, D. & De Maeyer, M. (1999). Guiding a docking mode by phage display: selection of correlated mutations at the staphylokinase-plasmin interface. *J. Mol. Biol.* **290**, 471-479.

Jones, S. A., Dewald, B., Clark-Lewis, I. & Baggiolini, M. (1997). Chemokine antagonists that discriminate between interleukin-8 receptors. Selective blockers of CXCR2. *J. Biol. Chem.* **272**, 16166-16169.

Kelner, C. S., Kennedy, J., Bacon, K. B., Kleyensteuber, S., Largaespada, D. A., Jenkins, N. A., Copeland, N. G., Bazan, J. F., Moore, K. W., Schall, T. J. & Al, E. (1994). Lymphotactin: a cytokine that represents a new class of chemokine. *Science,* **266**, 1395-1399.

Kim, C. H. & Broxmeyer, H. E. (1999). Chemokines: signal lamps for trafficking of T and B cells for development and effector function. *J. Leukocyte Biol.* **65**, 6-15.

Liao, F., Alkhatib, G., Peden, K. W., Sharma, G., Berger, E. A. & Farber, J. M. (1997). STRL33, A novel chemokine receptor-like protein, functions as a fusion cofactor for both macrophage-tropic and T cell line-tropic HIV-1. *J. Exp. Med.* **185**, 2015-2023.

Locati, M. & Murphy, P. M. (1999). Chemokines and chemokine receptors: biology and clinical relevance in inflammation and AIDS. *Annu. Rev. Med.* **50**, 425-440.

Lu, B., Humbles, A., Bota, D., Gerard, C., Moser, B., Soler, D., Luster, A. D. & Gerard, N. P. (1999). Structure and function of the murine chemokine receptor CXCR3. *Eur. J. Immunol.* **29**, 3804-3812.

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

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature,* **402**, 83-86.

Moyle, W. R., Campbell, R. K., Myers, R. V., Bernard, M. P., Han, Y. & Wang, X. (1994). Co-evolution of ligand-receptor pairs. *Nature,* **368**, 251-255.

Oppenheim, J. J., Zachariae, C. O., Mukaida, N. & Matsushima, K. (1991). Properties of the novel proinflammatory supergene intercrine cytokine family. *Annu. Rev. Immunol.* **9**, 617-648.

Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523.

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

Petersen, F., Brandt, E., Lindahl, U. & Spillmann, D. (1999). Characterization of a neutrophil cell surface glycosaminoglycan that mediates binding of platelet factor 4. *J. Biol. Chem.* **274**, 12376-12382.

Premack, B. A. & Schall, T. J. (1996). Chemokine receptors: gateways to inflammation and infection. *Nature Med.* **2**, 1174-1178.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK.

Rollins, B. J. (1997). Chemokines. *Blood,* **90**, 909-928.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425.

Tensen, C. P., Flier, J., Van Der, Raaij-Helmer E. M., Sampat-Sardjoepersad, S., Van Der, Schors R. C., Leurs, R., Scheper, R. I., Boorsma, D. M. & Willemze, R. (1999). Human IP-9: a keratinocyte-derived high affinity CXC-chemokine ligand for the IP-10/Mig receptor (CXCR3). *J. Invest. Dermatol.* **112**, 716-722.

Thompson, J. D., Higgins, D. C. & 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. *Nucl. Acids Res.* **22**, 4673-4680.

Wang, J. M., Deng, X., Gong, W. & Su, S. (1998). Chemokines and their role in tumor growth and metastasis. *J. Immunol. Methods,* **220**, 1-17.

Zaballos, A., Gutiérrez, J., Varona, R., Ardavín, C. & Márquez, G. (1999). Cutting edge: identification of the orphan chemokine receptor GPR-9-6 as CCR9, the receptor for the chemokine TECK. *J. Immunol.* **162**, 5671-5775.

Edited by B. Honig

(Received 20 December 1999; received in revised form 27 March 2000; accepted 28 March 2000)