Investigating the evolution and structure of chemokine receptors

Pietro Lio\textsuperscript{a,b,*}, Marina Vannucci\textsuperscript{c}

\textsuperscript{a}Department of Zoology, University of Cambridge, Cambridge, UK
\textsuperscript{b}EMBL-EBI, Hinxton, Cambs, UK
\textsuperscript{c}Department of Statistics, Texas A&M University, USA

Received 1 August 2002; received in revised form 16 October 2002; accepted 29 April 2003

Abstract

Chemokine receptors represent a prime target for the development of novel therapeutic strategies in a variety of disease processes, including inflammation, allergy and neoplasia. Here we use maximum likelihood methods and bootstrap methods to investigate both the phylogenetic relationships in a large set of human chemokine receptor sequences and the relationships between chemokine receptors and their nearest neighbors. We found that CCR and CXCR families are not homogeneous. We also provide evidences that angiotensin receptors are the closest neighbors. Other close neighbors include opioid, somatostatin and melanin-concentrating hormone receptors. The phylogenetic analysis suggests ancient paralogous relationships and establishes a link between immune, metabolic and neural systems modulation. We complement our findings with a structural analysis based on wavelet methods of the major branches of chemokine receptors phylogeny. We hypothesize that receptors very close in the tree can form heterodimers. Our analyses reveal different characteristics of amino acid hydrophobicity and volume propensity in the different subfamilies. We also found that the second extra-cytoplasmic loop has higher rates of evolution than the internal loops and transmembrane segments, suggesting that selection, shifting, reassignments and broadening of receptor binding specificities involve mainly this loop.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Phylogeny; Chemokine receptors; Angiotensin receptors; Somatostatin receptors; Opioid receptors

1. Introduction

Chemokine receptors (CR) are G-coupled protein receptors that bind chemokines (chemotactic cytochines) that are small peptides. Chemokines and their receptors are essential components of hematopoiesis, leukocyte trafficking, organogenesis and immuno-modulation in mammalian species (Baggiolini, 1998). Chemokines are involved in a variety of disease processes including inflammation, allergy and neoplasia. In particular, chemokines are key regulators of innate and adaptive immune responses against invading microorganisms, including viruses. All chemokines have four conserved cysteines linked to disulfide bonds. The positioning of two conserved cysteines in the N-terminal region of the proteins defines four chemokine subfamilies: CXC, CC, C, and CX3C (Rollins, 1997; Luster, 1998). To date a number of human receptors specific for these chemokine subfamilies have been described, though many receptors are still unassigned. Several viruses, for example, Epstein–Barr, Cytomegalovirus, and Herpes Samiri, contain functional homologues to human CRs, an indication that such viruses may use these receptors to subvert the effects of host chemokines (Murphy, 2001). Moreover, primate immuno-deficiency retroviruses have adopted CRs as essential gateways for entry into their target cells (Owen et al., 2000).

Few authors have investigated the evolution of CRs (Shields, 2000; Goh et al., 2000). These authors only indirectly addressed the problem of the phylogeny of CRs and they did not use robust methods such as maximum likelihood or different models of evolution and site heterogeneity of mutation rates. Little is also known about the evolutionary relationship between CRs and other G-coupled receptor proteins. This knowledge would help understand the evolution of immuno modulation signalling with respect to other signalling systems. Therefore, we present here the results of the analysis of a large set of human CR sequences and we investigate the nearest neighbor sequences. We found that the phylogenetic relationships between CRs and their neighbor G-coupled receptors also reflect the functional relatedness between these proteins.
Despite the importance of chemokines and chemokine receptors in medicine, there is also a lack of structural data. The main reason for this paucity of studies is the difficulty of determining the structure of membrane proteins, using, for example, nuclear magnetic resonance spectroscopy or X-ray spectroscopy, because G-coupled receptors are generally very large and do not crystallize.

Research into both molecular evolution and protein structure has much to gain in the post-genomic era. However, to make optimal use of the genomic and proteomic data that are now becoming available, it is necessary to exploit the relationships between evolution and structure. In other words, since proteins are a result of evolution, it is crucial to understand the process of evolution in order to understand the properties of proteins. Mutations, in fact, erode structural similarity at a much slower rate than sequence similarity. Fundamental to this approach must be the representation of the evolution of sequences along lineages of evolutionary trees, as these trees describe the complex patterns of dependence amongst sequences that are caused by their common ancestry (see Whelan et al., 2001 and references therein).

We have combined evolutionary and structural investigations by contrasting the site heterogeneity of rates of evolution with the structural organization of chemokine receptors. We found that the external loops of CRs have higher rates of evolution than internal loops and transmembrane segments. We used wavelet methods to analyze different structural properties of CRs, such as residue hydrophobicity and volume. Our analyses reveal structural similarities and differences within the subfamilies of CRs.

2. Materials and methods

Human CR sequences were obtained from GenBank and Swiss-Prot databases: CXCR1–CXCR5, CCR1–CCR10, XCR1 lymphotactin receptor, and CX3CR1 fractalkine (CX3C). We have also analyzed the human Duffy receptor (DARC), US28 (human cytomegalovirus–herpesvirus 5 receptor), EBI2 (Epstein–Barr virus induced G-protein coupled receptor), angiotensin receptors (AGTR1,2, APJ), somatostatin I-V receptors, kappa, delta and mu opioid receptors, melanin concentrating hormone receptor (MCHR 1,2), Mas1.

2.1. Phylogeny assessment

Because of the reliance on a strong statistical foundation and the complete utilization of the sequence information, likelihood is the most suitable method for phylogeny construction presently available (Whelan et al., 2001). Likelihood maximization and maximum likelihood (ML) parameter estimation were performed by numerical optimization routines. All likelihood calculations used standard methods described, for example, by Swofford et al. (1996). The assessment of phylogenies using likelihood framework depends on the choice of an evolutionary model. We computed the maximum likelihood (ML) analysis of the CRs data set under different models of evolution: Dayhoff (Dayhoff et al., 1978), JTT (Jones et al., 1992), WAG (Whelan and Goldman, 2001), THEM (Li and Goldman, 1999), and TM, which is a substitution matrix derived from transmembrane segments (Li and Goldman, 1999). The performance of the TM model gives an idea of how much importance transmembrane regions have in the likelihood calculation. We used these models considering the incorporation of the amino acid frequencies of the chemokine data sets, (+F), and the heterogeneity of the rates of evolution, implemented using a gamma distribution (+G; Yang, 1994, 1997). Bootstrap and permutation tests have been used to assess the robustness of the tree topology (Whelan et al., 2001).

2.2. Structural analysis

We have used wavelets to analyze the variance of structural profiles of CRs. We considered two structural parameters: hydrophobicity (Kyte and Doolittle, 1982) and volume (increase in volume of water after adding 1 g of residue expressed in cubic centimeters per gram; Zamyatin, 1972). In the last few years, wavelets (Daubechies, 1992) have emerged as a powerful tool in protein structural investigations: primary sequence evolution (Morozov et al., 2000; Rzhetsky and Morozov, 2001), secondary (Li and Vannucci, 2000; Vannucci and Liò, 2001) and tertiary structure determination (Murray et al., 2002; Mandell et al., 1998; Hirakawa et al., 1999), refinement of X-ray crystallography (Main and Wilson, 2000; Ferrer et al., 1998), drug design and visualization (Carson, 1996). In particular, Liò and Vannucci (2000) applied wavelet smoothing techniques to transmembrane proteins and obtained a per-transmembrane segment accuracy comparable with other currently used prediction methods. Recently, Murray et al. (2002) used wavelets to analyze the hydrophobicity and relative accessible surface area of a variety of repeating protein motifs such as TIM barrels, propeller blades, coiled coils and leucine repeats.

Wavelets are approximating functions that oscillate like waves and that decay rapidly to zero outside a finite interval. A wavelet transform (Mallat, 1989) decomposes a signal into several groups of coefficients. Coefficient sets at different levels contain information about characteristics of the data sequence at different scales or sequence-lengths. The scale parameter in the wavelet analysis has an analogy with maps; coarse scales correspond to a non-detailed global view of the signal and fine scales correspond to a detailed view. A useful interpretation of a wavelet transform is as a cumulative measure of the variations in the data over regions proportional to the wavelet scales, with coefficients at coarser and coarser levels describing features at lower frequency ranges and larger sequence lengths. In the non-decimated transform, features at different scales are aligned with those of the original data sequence. An inverse wavelet
transform can be defined and allows reconstruction of a signal from its wavelet decomposition (Mallat, 1989).

We used wavelets to estimate the variance of profiles containing structural information (Percival, 1995). Replacing 'global' variability with variability over scales allows us to investigate the effects of selection constraints acting at different sequence scales. An estimate of the wavelet variance at a given scale is obtained by summing the squares of the wavelet coefficients and dividing by the number of them.

To illustrate the intuition behind our work, we have reproduced in Fig. 1 three “virtual” profiles of ideal transmembrane proteins, one with equally spaced transmembrane helices of same width of 20 bp (a), one with irregularly spaced helices with same width (b) and one with irregularly spaced helices with different widths (c). The wavelet variance decompositions of the three profiles are shown in plots (d)–(f), respectively. All levels with wavelet coefficients are represented. The variance values are represented by “+” while the circles indicate the extremes of the 95% confidence interval. We grouped variances of CRs using boxplots. A boxplot is a graphical tool that allows to look at the overall shape and variability of a set of data, in our case wavelet variances, at different levels, of profiles of chemokine receptors belonging to the same family. A box shows the data between the 25th and 75th percentiles, that is the “central” portion of the data. The median is represented by a white horizontal line. Dotted vertical lines go out to the extremes of the data and very extreme data points (“outliers” in statistical terminology) are shown as isolated points. All profiles show large variance values at levels 5, capturing changes in averages of the data on a scale of $2^5=16$ bp. Less regularly spaced helices result in large values also at level 6, that is at a scale of $2^6=32$ bp. Clearly, structural diversities at various sequence-lengths among profiles are reflected in differences in the corresponding wavelet variances at the different scales. In all analyses of this paper we have used Daubechies (1992) wavelets. There are many families of these wavelets, essentially characterized by different smoothness and regularity. A limited exploratory analysis we performed resulted in the choice of minimum phase wavelets with two vanishing moments (see Daubechies, 1992, Chap. 6, pp. 194–197, for a description and plots of this family). Matlab (www.mathworks.com) programs developed for the analyses of this paper use the Wavbox package (Toolsmith by Carl Taswell www.wavbox.com) and are available upon request from the authors.

3. Results

3.1. Phylogenetic relationships of chemokine receptors

Despite their importance in medicine, only a few authors have investigated the evolution of chemokine receptors. Their studies were based on neighbor-joining (Shields, 2000) or correlation methods (Goh et al., 2000). In our study

![Fig. 1. Three signals resembling ideal transmembrane proteins, one with equally spaced transmembrane helices of the same width (a), one with irregularly spaced helices with same width (b), one with irregularly spaced helices with different widths (c) and their corresponding variance decompositions (d–f).](image-url)
we have used more statistically robust methods, such as maximum likelihood, bootstrap and permutation methods, and analyzed a larger set of human CRs than in previous studies, taking into account the CRs binding all classes of chemokines. We explored a certain number of available models of evolution (Dayhoff, Jones, Wag, THEM, TM) and statistical assumptions, i.e. the incorporation of amino acid frequencies from the data and using a gamma distribution to model the heterogeneity of rates at different sequence sites. We found that all models of evolution suggest the same tree topology, with all CRs divided into four major branches. The JTT+F+C model gave the highest ML value (Table 1); WAG and JTT gave very similar results (Whelan and Goldman, 2001). The ML tree, obtained using the JTT+F+C model of evolution, is shown in Fig. 2, plot (a). A bootstrap analysis with 1000 replicates gave a statistically robust support (>90%) for the major branchings of the tree. The topology clearly shows, in contrast to previous studies, that the CCR family is not homogeneous: CCR6, CCR7, CCR9 and CCR10 are separated from the other CCRs; in particular, CCR10 clusters with CXCRs (branch 3); CXCR4 and CXCR6 do not cluster with the CXCRs. Note that angiotensin II type I and somatostatin IV receptors cluster with a G-coupled protein receptor induced by the Epstein–Barr virus (EB12). The distances between the sequences in branches 1–3 and the angiotensin/somatostatin branch are almost the same. The viral protein US28 from cytomegalovirus is close to CX3CR.

3.2. Contrasting site heterogeneity of mutation rates and structural organization

Table 1 shows that, for all models of evolution, modeling mutation rates using a gamma function ‘+F’ always lead to larger improvement in ML values with respect to models that assume constant rates of evolution. This suggests that there is a large heterogeneity of rates among sites in the CR sequences. The improvement in ML also occurs when using the amino acid frequencies of the CR data set (‘+F’). Using the ML tree as a framework, we have analyzed the heterogeneity
of rates among sites in the three major branches (1–3) of the tree (Fig. 2, plot a). First, we have multialigned the sequences in each of the three branches and performed a ML-based phylogenetic analysis of the three sets of sequences using the JTT+F+F+I model. The calculation of the ‘a posteriori’ probabilities allows to estimate the sitewise rates of evolution (Yang, 1997). Then we have used wavelets to compute the variance of the mutation rates at different sequence-lengths and compared with transmembrane locations. Variance values are shown in Fig. 2, plot (b). For clarity we do not report the 95% confidence intervals. The largest values occur at scale 6 for all three profiles, indicating that there are predominant features of the signals at such scale. These features can be extracted by applying an inverse wavelet transform that uses only the coefficients at level 6, essentially replacing by zeros all the other coefficients. Fig. 2, plot (c), shows the reconstruction of the heterogeneity profile of the CCRN set in branch ‘1’ at scale 6 (line) and the positions of the transmembrane segments of CCR5 (from Swiss-Prot). Note that the peaks of the reconstructed heterogeneity profile correspond to the extracellular loops and to the N- and C-terminal domains of the CCR5 sequence. Sequences in branch 1 have very high percentage of amino acid sequence identity (75%); therefore they have similar transmembrane segments locations. Our conclusions from this analysis are that the residues of the internal loops, involved in signal transducing, are more conserved than residues in extracellular loops, involved in chemokine binding specificity. In particular, the second extracellular loop shows the largest rate heterogeneity. This finding is in agreement with the discovery of Samson et al. (1997) that the second extracellular loop of CCR5 is the major determinant of ligand specificity. We found a similar result for the sequences in branches 2 and 3.

3.3. Investigating the structural organization of CR

The transmembrane nature of the receptor impedes the structure-based understanding of ligand interactions. This explains the lack of structural data about chemokine receptors in the Brookhaven PDB database. In our work we have investigated the structural properties of CRs by comparing the variances, computed using wavelets, of different subfamilies of chemokines. In particular, we consider amino acid hydrophobicity and volume. It is recognized that hydrophobicity is the most important property guiding membrane protein folding and stability (Popot and Engelman, 1990), and that it is the most important property affecting protein evolution (Xia and Li, 1998). The volume of the residues is also known to affect protein evolution (Pollock et al., 1999).

We have used the information extracted from the evolutionary inference analysis to investigate the structural properties of sequences in the three major branches of the tree in Fig. 2 and report here on hydrophobicity and volume. Fig. 3 shows the boxplots of wavelet variances of the hydrophobic-
ity profiles of data sets of branches 1–3 (a–c, respectively), and those of volume profiles (d–f, respectively). While the sequences in branch 1 show large differences, in particular at scale 6, the sequences in branch 2 seem to be more homogeneous. The sequences in the three branches show larger differences in volume; again sequences of branch 1 are less homogeneous and have larger values at level 6. Wavelet variances for single member of the CCRs and CXCRs are not reported here but can be made available by the authors.

In order to statistically validate the information provided by the wavelet variance analysis, we performed a permutation test in the following way. In each of the branches 1 and 3 we computed the means and standard deviations of the wavelet variance values, at the different scales. We used the ratio of the mean difference to the square root of the pooled variance as test statistic for a null hypothesis of no difference in distribution between the two branches. The permutation distribution under the null hypothesis was used to obtain p-values at the different scales. The test is performed by computing all possible permutations of the sequences in the two branches and counting the number of times that the test statistic ratio exceeds the value observed on the original tree. Scale 6 resulted statistically significant with a p-value of 0.019, confirming the observed difference, at that scale, between sequences in the two branches. We also got a p-value of 0.024 at scale 6 for branches 1 and 2.

3.4. Nearest neighborhood of chemokine receptors

A Blast search (Altschul et al., 1990) revealed that CCR5 is more similar to angiotensin II type I receptor (31% amino acid identity; 50% amino acid similarity) and to somatostatin IV receptor (29% identity; 49% similarity) than to other CRs, such as CCR10 (30% identity; 47% similarity). We found that chemokine receptors sequences share large sequence homology with angiotensin, somatostatin, melanin-concentrating hormone and opioid receptors. A Blast search with a query sequence belonging to any of these four protein families gave as results sequences also from the other families. Sequences from other families have much lower p-values. We obtained equal results using Psi-Blast (Altschul et al., 1997).

While pairwise comparison of sequences can be misleading, phylogenetic methodology allows to recognise and exploit the statistical dependencies among sequences sharing a common ancestry. For example, there are cases in which protein functions are inferred from pairwise sequence similarities to proteins whose functions have themselves been inferred from pair-wise sequence similarities. Since paralogous proteins may not have the same function, the phylogenetic relationships among sequences allow each species to contribute its own quantity to all estimates. Phylogenetic methods also make appropriate corrections for multiple hits. We investigated CRs’ nearest neighbours, by using ML methods and a certain number of available models of evolution (Dayhoff, Jones, Wag, THEM, M) and statistical assumptions (F+I). We found that all models of evolution suggested the same tree topology and the WAG+F+I model gave the highest ML value (Table 2). Fig. 4 shows the maximum likelihood phylogeny under the
4. Discussion

In this paper we have analyzed the structural and evolutionary characteristics of CRs using wavelet and phylogenetic methods. Phylogenetic analyses show that CRs do not cluster precisely according to the four chemokine subfamilies (CC, C, CXC, CX3C). In particular the CCR family is split into two groups; CXCR4 and CXCR6 are distant from the other CXCRs. The difference between the observed topology and the four class divisions of CRs may depend on reassignment or broadening of the receptor binding specificities during evolution. There are several clues supporting this hypothesis. For instance, CXCR6 (also termed Bonzo, strl33, tymstr) also has the characteristics of CCRs (Wilbanks et al., 2001). DARC, located isolated in the topology, is known to bind both CXC and CC chemokines. Interestingly, HIV uses most of branch 1 CCR as co-receptors. In particular the distances between CCR5, CXCR3 and CXCR4 may reflect the existence of specific tropism of HIV strains: T-tropic strains which selectively interact with the CXCR4 to infect lymphocytes; the M-tropic strains of HIV interact with the CCR5 chemokine co-receptor to infect macrophages. Dual tropic HIV strains have also been identified. In many cases, the binding specificity involves one or few residues. Recently, Speck et al. (1997) reported that individual amino acids within the envelope V3 loop determine the selective employment of CCR5 and CXCR4. The fractalkine binds both US28 and cx3cr (Combadiere et al., 1998); analysis of the mutants indicates that US28 recognizes many of the same epitopes of fractalkine as CX3CR1 (Mizoue et al., 2001). We found that Eb2, which was identified by the up-regulation of its expression upon Epstein–Barr virus infection of primary B lymphocytes, is close to chemokine receptors. The angiotensin receptors are the closest neighbours to CRs; AGTR1 and AGTR2 bind the vasopressor angiotensin II that is an important effector controlling blood pressure and volume in the cardiovascular system (AbdAlla et al., 2000). Interestingly, these angiotensin receptor genes are located in the same chromosomal regions of several CCR genes.

Our results suggest that Mas1 may be related to the angiotensin receptors. This oncogene is related to the large family of approximately 50 GPCRs called mrg (Dong et al., 2001) that may regulate nociceptor function and/or development, including the sensation or modulation of pain. Other close neighbours to CRs include opioid, somatostatin and melanin concentrating hormone receptors.

Note that the tree topology also reflects some functional similarities between CRs and their neighbor receptors. For example, Cayabyab et al. (2000) reported that primary HIV-1 isolates can also use the angiotensin-like APJ. The opioid system modulates several physiological processes, including analgesia, the stress response, and neuroendocrine function. Recent works show it also modulates the immune response (Jordan and Devi, 1999); the immune modulation by the opioid receptors may trace back to the ancient paralogy with chemokine receptors. The hormone somatostatin inhibits the release of many physiologically important compounds, including growth hormone, insulin, glucagon, gastrin and secretin. It is therefore a major factor in the control of metabolism. It also functions in the central nervous system as a neurotransmitter/neuromodulator in the control of motor activity and cognitive functions (Rohrer et al., 1993). Note that MCH receptors, which are closer to somatostatin receptors, are involved in the regulation of feeding (Qu et al., 1996). Therefore, these phylogenetic relationships suggest ancient paralogous relationships and establish a link between immune, metabolic and neuro systems modulation.

4.1. Phylogeny and heterodimerization properties of chemokine receptors

Little is known about the ability of the chemokine receptors to interact in order to form new functional structures, the simplest of which would be a dimer. Chemokine receptors of both the CC and CXC families have shown to undergo a ligand-mediated homodimerization process. Recent work shows that heterodimerization is also permitted between given receptor pairs, for example, between CCR2 and CCR5. This has functional consequences, as the CCR2 and CCR5 ligands, MCP-1 and RANTES cooperate to trigger calcium responses at concentrations 10- to 100-fold lower than the threshold for either chemokine alone. Heterodimerization results in the recruitment of each receptor-associated signaling complex, but also recruits dissimilar signaling pathways (Mellado et al., 2001). Therefore, heterodimerization results in a new receptor that exhibits ligand binding and functional properties that are distinct from those of either receptors. Note that CCR2 and CCR5, and CXCR1 and CXCR2, for which there is structural and biochemical evidence that they form heterodimers, are very close in the tree. There are biochemical and pharmacological evidence for the heterodimerization of the kappa and delta opioid receptors. Heterodimerization of these receptors represents a novel mechanism that modulates their function. Note that heterodimerization requires domain conservation in two different receptors or their coevolution. This will result in a close position in the tree. In particular, it is suggested that the correlated residues may be involved in the formation of heterodimers or homodimers.

4.2. Structural organization

We have found that rates of evolution are larger in external loops than in internal loops and transmembrane
segments. In particular the second extracytoplasmic loop has the largest variation. Larger volume values at the smallest time-scales show that, in general, large amino acids do not cluster, suggesting that there is a certain selection against runs of large residues. This is in agreement with the idea that large residues with more conformational flexibility are disfavored for entropic reasons. During the folding of the membrane protein, large side chains may experience a steep loss of conformational entropy (see for instance Koshi and Goldstein, 2001 and references therein). Thus, runs of large side chains with more conformational entropy would have a greater contribution to this effect opposing folding, generally disfavoring clustering of large residues at all positions. The sequences in branch 1 show larger values at level 6 than the sequences in branch 2, suggesting that in these sequences large residues are rare or absent in short loops.

The aim of this paper has been to investigate chemokine phylogeny and to increase the familiarity with chemokine evolution in the medical bioinformatics community. In conclusion, we are delighted to report that the investigation of chemokine evolution is a thriving field of research. It has two important and linked benefits: it captures information that can be valuable in pharmacology, investigating the links between immune, endocrine and neuro systems hidden components from biological data, and it provides a better link between biological systems and the mathematics objects used to describe them.

Acknowledgements

P. Liò is supported by an EPSRC/BBSRC Bioinformatics Initiative grant. M. Vannucci is supported by National Science Foundation, CAREER award number DMS-0093208. We thank Nick Goldman for helpful suggestions.

References


