Isolation of Novel GRO Genes and a Phylogenetic Analysis of the CXC Chemokine Subfamily in Mammals

William S. Modi* and Teizo Yoshimura†

*Intramural Research Support Program, SAIC Frederick, and †Immunopathology Section, Laboratory of Immunobiology, National Cancer Institute-Frederick Cancer Research and Development Center, Frederick, Maryland

Approximately 15 different α, or CXC, chemokines have thus far been isolated from 11 species of mammals. Among the best studied chemokines are the 12 human proteins that are encoded by 11 paralogous genes. In order to better understand the evolution and function of this group of genes, we isolated and characterized six novel GRO and GRO-related cDNA sequences from the cow (Bos taurus), the sheep (Ovis aries), the rabbit (Oryctolagus cuniculus), and the guinea pig (Cavia porcellus). The amino acid sequence of the diverged guinea pig GRO or KC gene is only 50%–60% similar to presumed orthologs from other species, while the sheep and cow GRO proteins are 90%–99% similar to each other. The presence of multiple GRO genes in the cow, the rabbit, and the sheep is consistent with what has been observed for humans. Phylogenetic analyses of amino acid sequences from 44 proteins indicate that genes orthologous to many of the 11 known from humans exist in other species. One such gene, interleukin 8, or IL8, has been isolated from nine species, including the rodent guinea pig; however, this gene is absent in the rat and the mouse, indicating a unique gene loss event in the rat/mouse (muroid rodent) lineage. The KC (or MIP2) gene of rodents appears to be orthologous to the GRO gene found in other taxonomic orders. Combined evidence from different sources suggests that IP10 and MIG share sister taxon relationships on the evolutionary tree, while the remaining paralogous genes represent independent lineages, with limited evidence for kinship between them. This observation indicates that these genes originated nearly contemporaneously via a series of gene duplication events. Relative-rate tests for synonymous and nonsynonymous nucleotide substitutions in the KC and IL8 genes did not detect rate heterogeneity; however, there are several notable features regarding the IL8 genes. For example, the IL8 proteins from two Old World monkeys are as similar to one another as they are to the IL8 protein from humans, and all observed nucleotide differences between the IL8 genes of the two monkeys cause amino acid changes; in other words, there are no synonymous differences between them.

Introduction

Over the past 2 decades, many members of a new class of chemotactic cytokines have been identified. These chemokines are small (8–10 kDa) basic heparin-binding proteins that exhibit 25%–80% amino acid similarity with each other. The amino acid sequences contain four conserved cysteine residues, and these molecules have various proinflammatory and reparative activities (Oppenheim et al. 1991). The specific effects of chemokines on the appropriate target cells are mediated by receptors belonging to the G-protein-coupled seven-transmembrane domain receptor family (see Baggiolini, DeWald, and Moser 1994). Chemokines are typically divided into two subgroups. The α, or CXC, subgroup is characterized by an intervening amino acid present between the first two cysteine residues, whereas this intervening residue is lacking in the β, or CC, subfamily members.

Among the best studied CXC chemokines are the 12 proteins encoded by 11 genes that have currently been identified from the human genome (see table 1); however, the exact biological functions of most of them are not precisely known. Although their chemoattractive effects on neutrophils have been well documented, the manner in which the different proteins are created and the cell types from which they are secreted differ markedly, suggesting that their functions may be much more complex. For example, interleukin 8 (IL8) is produced and secreted by monocytes and other cell types in response to lipopolysaccharide (LPS), tumor necrosis factor (TNF), or IL1 stimulation (see Baggiolini, DeWald, and Moser 1994). Alternatively, pro-platelet basic protein (PPBP) is derived from inactive precursors released from platelet granules by proteolytic cleavage, and this process occurs in the presence of monocytes or their supernatants (Walz and Baggiolini 1990). Thus, PPBP probably functions in the vasculature, where platelet activation may occur along with inflammation (Walz 1992). Platelet factor 4 (PF4) is a platelet-specific protein that is synthesized in megakaryocytes and packaged into the platelet α-granules (Handin and Cohen 1976). This protein may neutralize heparin-like molecules in the serum and on endothelial cell surfaces, thus decreasing thrombin III activity and allowing coagulation to proceed (Marcum, McKenney, and Rosenberg 1984). Further, the in vitro activity of PF4 on neutrophils is considerably lower than that of either IL8, GRO, or PPBP (Deuel et al. 1981; Walz 1992). Finally, in addition to possessing chemoattractant properties, SDF1 was also shown to be a pre-B-cell growth factor in the presence of IL7 (Nagasawa et al. 1994).

One GRO gene was originally identified by its constitutive overexpression in a highly tumorigenic Chinese hamster (Cricetulus grisescus) cell line (Anisowicz, Bardwell, and Sager 1987). Richmond et al. (1988) independently purified a protein that stimulated the growth of melanoma cells, termed melanoma growth stimulating activity (MGSA). The MGSA cDNA was subsequently cloned, and it was found that MGSA was iden-
Phylogeny of Mammalian CXC Chemokines

Table 1
Sequences Analyzed in the Present Study

<table>
<thead>
<tr>
<th>Gene Symbol, Protein Name, Species</th>
<th>Accession Number(s)</th>
<th>Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>J03561</td>
<td>107</td>
</tr>
<tr>
<td>Human</td>
<td>M36820</td>
<td>107</td>
</tr>
<tr>
<td>Human</td>
<td>M36821</td>
<td>106</td>
</tr>
<tr>
<td>Cow α</td>
<td>U95812</td>
<td>104</td>
</tr>
<tr>
<td>Cow β</td>
<td>U95813</td>
<td>104</td>
</tr>
<tr>
<td>Cow γ</td>
<td>U95811</td>
<td>98</td>
</tr>
<tr>
<td>Rabbit α</td>
<td>U95808</td>
<td>108</td>
</tr>
<tr>
<td>Rabbit β</td>
<td>L19157</td>
<td></td>
</tr>
<tr>
<td>Sheep</td>
<td>U95814</td>
<td>103</td>
</tr>
<tr>
<td>Human</td>
<td>M25897</td>
<td>104</td>
</tr>
<tr>
<td>Pig</td>
<td>P30034</td>
<td></td>
</tr>
<tr>
<td>Sheep</td>
<td>P30035</td>
<td></td>
</tr>
<tr>
<td>Cow</td>
<td>P02777</td>
<td></td>
</tr>
<tr>
<td>Rat</td>
<td>M15254</td>
<td>105</td>
</tr>
<tr>
<td>Human</td>
<td>X02530, M17752</td>
<td>98</td>
</tr>
<tr>
<td>Mouse</td>
<td>M33266, M37136</td>
<td>98</td>
</tr>
<tr>
<td>Rat</td>
<td>U22520</td>
<td>98</td>
</tr>
<tr>
<td>Human</td>
<td>L36033</td>
<td>89</td>
</tr>
<tr>
<td>Human</td>
<td>L36033</td>
<td>93</td>
</tr>
<tr>
<td>Mouse</td>
<td>L12029</td>
<td>89</td>
</tr>
<tr>
<td>Mouse</td>
<td>L12030</td>
<td>93</td>
</tr>
<tr>
<td>Human</td>
<td>X72755, S60728</td>
<td>125</td>
</tr>
<tr>
<td>Mouse</td>
<td>M34815</td>
<td>126</td>
</tr>
<tr>
<td>Mouse</td>
<td>X53798</td>
<td>100</td>
</tr>
<tr>
<td>Rat</td>
<td>U45965</td>
<td>100</td>
</tr>
<tr>
<td>Human</td>
<td>M54995, M38441</td>
<td>128</td>
</tr>
<tr>
<td>Pig</td>
<td>X77935</td>
<td>119</td>
</tr>
<tr>
<td>Human</td>
<td>P80162</td>
<td></td>
</tr>
<tr>
<td>Cow</td>
<td>P80221</td>
<td></td>
</tr>
</tbody>
</table>
| CINC2—cytokine-induced neutrophil chemoattractant
Rat α                            | D21095              | 100        |
Rat β                            | D21095              | 100        |
AMCFII—alveolar macrophage-derived chemotactic factor II
Pig                              | M99368              | 117        |
LIX—LPS-induced C-X-C chemokine LIX precursor
Mouse                            | U72767              | 132        |
ENA78—epithelial cell-derived neutrophil-activating peptide (also called small inducible cytokine subfamily B, member 5 [SCYB5])
Human                            | X78686              | 114        |

* The coding region contains the signal peptide and the mature peptide.
* Sequences contributed by the present study.
* For several proteins, the complete N-terminus has not yet been determined.
* The SDF1 and CINC2 α and β genes differ from one another only in the presence of three additional amino acids at the C-terminus of the β gene. As a result, the α genes are excluded from the alignment and analysis shown in figure 2.

In order to better understand the functions and evolutionary histories of this multigene family, the aims of the present study were threefold: (1) to isolate and characterize novel GRO and GRO-related cDNA clones from rabbits, cows, sheep and guinea pigs, (2) to conduct a phylogenetic analysis of all available CXC chemokine sequences in order to estimate which sequences from other species of mammals are orthologous to the 11 well-known human genes, and (3) to evaluate the relationships among paralogous gene lineages.

Materials and Methods

cDNA Library Construction

The construction of cDNA libraries from ConA-stimulated sheep, guinea pig, and rabbit spleen cells were previously reported (Yoshimura and Yukhi 1991; Yoshimura and Johnson 1993; Seow et al. 1994). A guinea pig macrophage cDNA library was prepared from strain 2 guinea pig resident peritoneal macrophages stimulated with 10 μg/ml of LPS (Escherichia coli...
cDNA and Genomic Library Screening

Approximately $5 \times 10^5$ recombinant phages from the guinea pig spleen cDNA library were screened by high-density plaque hybridization with a $^{32}$P-labeled oligonucleotide probe. The 30-bp reverse complement oligo, 5'-GTG GCT ATG ACT TCG GTT TGG GTG CAG TGG-3', is a consensus based on the nucleotide sequences of human GROβ, hamster KC, and mouse KC and covers portions of exons II and III. Hybridization to nitrocellulose filters was carried out overnight at $37^\circ C$ in a solution containing 6 × SSC, 5 × Denhardt’s solution, 0.05% sodium pyrophosphate, 1% SDS, 100 μg/ml salmon sperm DNA, and 1 × 10⁶ dpm probe. Filters were washed twice with 2 × SSC, 0.1% SDS at room temperature for 15 min and once at 50°C for 30 min, and were then dried and exposed overnight to RP-5 films (Kodak) with an intensifying screen at −80°C. Phagemids carried within a ZAP II recombinants were rescued with helper phage (Short et al. 1988). DNA sequencing was performed directly from double-stranded DNA by the dideoxynucleoside triphosphate chain termination method.

Approximately $5 \times 10^5$ individual clones from the cow, sheep, and rabbit cDNA libraries were each screened with a human GROβ cDNA clone. The hybridization solution contained 30% formamide, 5 × SSC, 5 × Denhardt’s solution, 1% SDS, 100 μg/ml salmon sperm DNA, and 5 × 10⁵ dpm/ml probe. Several positive clones from each library were selected for further restriction digestion and nucleotide sequence analysis.

To identify the 5’ end of the open reading frame of the guinea pig KC (or GRO) gene, a guinea pig genomic library (Stratagene, La Jolla, Calif.) was screened with a $^{32}$P-labeled guinea pig KC cDNA clone, termed 1A (described below). One genomic clone that contained a 3-kb EcoRI fragment was identified. The 3-kb insert was excised, subcloned into the phagemid vector pBluescript SK(−), and sequenced.

Southern Blot Analysis

Southern blot analyses were performed in 1% agarose gels with 10 μg restriction-enzyme-cleaved DNA per lane. Filters were hybridized at 42°C overnight in 50% formamide, 5 × SSC, 5 × Denhardt’s solution, 1% SDS, 100 μg/ml sheared-denatured salmon sperm DNA, and 1 × 10⁶ dpm/ml probe. Filters were washed twice with 2 × SSC, 0.5% SDS at room temperature for 15 min and with 0.1 × SSC, 0.5% SDS at 60°C for 60 min prior to autoradiographic exposure (Yoshimura and Johnson 1993).

Sequence Alignment and Phylogenetic Analyses

A total of 45 sequences were analyzed. Six of them are presented for the first time in the present study (GRO genes from the cow, the rabbit, and the sheep, and the KC gene from the guinea pig), while the remaining 39 were obtained from either the GenBank or the SwissProt database (table 1). The entire coding regions of all 45 proteins were aligned using PILEUP from the UWGCG package (Deveraux, Haberli, and Smithies 1984) and CLUSTAL W (Thompson, Higgins, and Gibson 1994). As a result of the low level of similarity between divergent proteins in the signal peptide or N-terminal region (left-hand side of fig. 2), this segment was excluded from the phylogenetic tree analyses. Further, since the extreme C-terminus also varied in sequence and length between proteins, this end was truncated so that the analyzed proteins would be approximately the same length. This resulted in the inclusion of a 66–70-amino-acid region of the mature peptide region in the evolutionary tree comparisons.

Evolutionary trees were constructed from the amino acid sequences using the neighbor-joining and maximum-likelihood algorithms with the computer programs NJDIST and PROTML from the MOLPHY package (Adachi and Hasegawa 1996).

Nucleic acid sequences from different species were aligned for each of the orthologous KC and IL8 genes in order to calculate nucleotide substitution rates. Sequences were aligned by PILEUP or CLUSTAL W and then adjusted by eye. The proportions of synonymous and nonsynonymous substitutions per synonymous and nonsynonymous site ($K_s$ and $K_a$) were calculated following the method of Li, Wu, and Luo (1985) using the DIVERGE program in the UWGCG package.

Results
cDNA Cloning of Guinea Pig KC

After screening approximately $5 \times 10^5$ ConA-stimulated spleen cell guinea pig cDNA clones with the oligo probe described above, a clone designated 1A was obtained. The nucleotide sequence of the insert showed a high similarity to the open reading frame sequence of the human GRO genes. Since the 5’ end of the open reading frame of the guinea pig gene was not included, an additional 2 × 10⁶ clones consisting of those from the original library and from an LPS-stimulated peritoneal macrophage cDNA library were further screened. Additional positive clones were obtained; however, none contained the 5’ end of the open reading frame. The sequence of the 5’ end of the coding region was subsequently obtained from a clone from a guinea pig genomic library.

The open reading frame of the guinea pig clone 1A cDNA encoded a 104-amino-acid protein (fig. 2) that showed 58%, 60%, and 59% amino acid sequence similarities to human GROα, GROβ, and GROγ sequences, 57% to mouse KC, 54% to mouse MIP2, 58% to rat KC, and 52% to rat MIP2.

Since humans are known to have three distinct GRO genes, while the rat and mouse have only one copy of each of two related genes known as KC and MIP2, it was of interest to estimate whether or not there are multiple copies of GRO-related genes in the guinea pig. This was ascertained by screening both guinea pig ge-
Fig. 1.—Southern blot hybridization of cDNA clones to restriction-enzyme-digested genomic DNA, illustrating the presence of one (A) or multiple (B–C) copies of KC/GRO in the genomes of the respective species: (A) guinea pig 1A cDNA clone hybridized to guinea pig DNA, (B) rabbit GROα clone hybridized to rabbit DNA, and (C) bovine 2F (GROα) probe hybridized to bovine genomic DNA.

nomic DNA using Southern blotting and the cDNA library with four different hybridization probes: the guinea pig 1A cDNA clone, a human GROβ cDNA clone, a rat KC cDNA clone, and a rat MIP2 cDNA clone. All hybridization analyses indicated that this sequence was present as a single copy in the guinea pig genome (fig. 1); however, the possibility remains that divergent copies of related genes exist in the guinea pig genome that were undetectable under the experimental conditions used here.

Characterization of GRO cDNAs in Rabbits, Cows, and Sheep

cDNA libraries from rabbits, cows, and sheep were screened using a human GROβ cDNA (Anisowicz et al. 1988) as a probe. A single clone containing a 975-bp insert, clone 14, was obtained after screening approximately 5 × 10⁵ phage clones from a rabbit cDNA library. The cDNA contained a 327-bp open reading frame that encoded a 108-amino-acid protein, and this sequence was designated rabbit GROα (fig. 2). Examination of GenBank revealed that there were at least three different types. The insert in clone 2F (GROα) was 1,046 bp in length, and the 315-bp open frame encoded a 104-amino-acid protein. Clone 2H (GROβ) contained a 363-bp insert and, similar to 2F, also encoded a 104-amino-acid protein. The insert in clone 2A (termed GROγ) contains 1,114 bp with a 297-bp open reading frame that encodes a 98-amino-acid protein. The open reading frames of clones 2A and 2F differ at 11 positions due to insertions, deletions, or substitutions, while those of 2F and 2H differ only due to two nucleotide substitutions. The length difference of six amino acids between 2A and the two other clones is attributable to the presence in clones 2F and 2H of three additional nucleotides (GCC) at positions 22–24 that encode the amino acid alanine, and to the presence at position 295 in clone 2A of a T that creates the termination codon TAG, whereas the corresponding position in clones 2F and 2H (position 298) contains an A, resulting in the amino acid lysine. The latter two clones continue an additional six codons beyond this lysine before the termination codon TGA is reached. In addition, clones 2F and 2H differ from 2A at six other amino acid positions, while the two nucleotide differences between 2F and 2H cause the single-amino-acid change of F to L (phenylalanine to leucine) at position 27 (fig. 2).

A single sheep GRO cDNA was obtained from a sheep cDNA library constructed from ConA-stimulated
FIG. 2.—Amino acid alignments of the complete coding regions (when available) for 45 α chemokine sequences from 11 species of mammals. The conserved cysteine (C) residues are located at positions 63, 65, 90, and 107. The functional ELR (glutamic acid, leucine, arginine) motif is located at positions 60–62. The amino acids used in the phylogenetic analysis range from positions 55 to 124. Gaps are denoted by periods. The last 10 amino acids for the two MIG proteins are excluded.

spleen cells. The deduced amino acid sequence was 90%–95% similar to the three bovine GRO proteins (see fig. 2).

Southern blot analysis was performed with rabbit, sheep, and bovine genomic DNAs using homologous GRO cDNA probes in each case. There appeared to be more than one copy of the GRO genes in each of these three species, as illustrated in figure 1 for the rabbit and cow. These hybridization data are consistent with results from cDNA cloning, which has identified three distinct genes in the cow and rabbit, while additional cloning is needed in order to isolate the remaining sheep genes.
At present, 11 human α, or CXC, chemokine genes are known, including SDF1, IL8, GROα, GROβ, GROγ, PF4, PPBP, MIG, IP10, ENA78, and GCP2. Two major questions are addressed here: (1) Which of the genes that have been isolated from other species are orthologous to the human genes, and (2) what are the relationships among the paralogous genes? In order to answer these questions, the 45 protein sequences listed in Table 1 were aligned and examined. The bovine GROα and GROβ proteins differ only at amino acid position 27, and since this residue was not included in the phylogenetic analysis, only 44 proteins were examined. All of the sequences examined contain the hallmark CXC motif (positions 63–65 in Fig. 2) and the two additional cysteines (positions 90 and 107). The following amino acids are conserved in many of the proteins: T (position 68), K (76), V (83), GPHC (87–90), EVIATL (94–99),...

Phylogenetic Analyses of the CXC Chemokines

At present, 11 human α, or CXC, chemokine genes are known, including SDF1, IL8, GROα, GROβ, GROγ, PF4, PPBP, MIG, IP10, ENA78, and GCP2. Two major questions are addressed here: (1) Which of the genes that have been isolated from other species are orthologous to the human genes, and (2) what are the relationships among the paralogous genes? In order to answer these questions, the 45 protein sequences listed in Table 1 were aligned and examined. The bovine GROα and GROβ proteins differ only at amino acid position 27, and since this residue was not included in the phylogenetic analysis, only 44 proteins were examined. All of the sequences examined contain the hallmark CXC motif (positions 63–65 in Fig. 2) and the two additional cysteines (positions 90 and 107). The following amino acids are conserved in many of the proteins: T (position 68), K (76), V (83), GPHC (87–90), EVIATL (94–99),...
CLD/NP (107–110), K (117), and L (123). Finally, the SDF1, PF4, MIG, and IP10 proteins all lack the characteristic ELR motif (positions 60–62) found among the other CXC chemokines.

The maximum-likelihood tree obtained using PROTML is shown in figure 3 and has a log likelihood of -2,783.88, while the log likelihood of the corresponding neighbor-joining tree was 9.769.8 lower than that of the maximum-likelihood tree. In this study, clustering relationships on the maximum-likelihood tree, overall amino acid similarity, and the presence or absence of conserved motifs are all used in estimating relationships among the different proteins.

**Comparisons Among Orthologous Genes**

In the case of the SDF1 proteins from humans and mice, these proteins are 98% identical to one another over the 70-amino-acid region examined here and are
Human-GROb KNGQKACLNP ASPMVKKIIE KMLNGKSN.
Human-GROg KNGKKACLNP ASPMVKKIIE KILNKGNST.
Rat-KC KNGREACLDP EAPMVQKIVQ KMLKGVPK.
Mouse-KC KNGREACLDP EAPLVQKIVQ KMLKGVPK.
Hamster-KC KNGQEACLNP EAPMVQKIVQ KMLSGIRK.
Mouse-MIP2 KGGQKVCLDP EAPLVQKIIQ KILNKGKAN.
Rat-MIP2 KDGHEVCLNP EAPLVQRIIVQ KILNKGKAN.
Rat-CINC2 KDGQEVCLNP QAPRLQKIIQ KLKSPSL.
Guinea pig-KC KGREACLDP EAPMVQKVLQ RMLKGSKAT.

Fig. 2 (Continued)

Fig. 3.—The unrooted maximum-likelihood tree of 44 chemokine proteins constructed using the PROTML program (JTT model).
almost certainly orthologous. SDF1 is the most diverged member of the CXC subfamily: the human protein shares only 20%–32% amino acid similarity to any of the paralogous human proteins.

The IL8 sequences from nine different species group together on the tree and share between 70% (dog [Canis familiaris] and guinea pig) to 94% (sheep and cow) amino acid similarity. These proteins are sufficiently conserved so that the entire coding regions of about 100 amino acids may be compared. The IL8 sequences are highly similar in the first 20 amino acids (the signal peptide region) and in the 40-amino-acid stretch following the conserved ELR motif. The final 30 amino acids of the protein toward the C-terminus is the most diverged region of the protein. The three primate sequences ally very close together. Interestingly, these three proteins (human and two Old World monkeys, mangabey [Cercocebus torquatus] and macaca [Macaca nemestrina]) are about equidistant from one another, differing at only five or six amino acids (fig. 2). Sequences from two (sheep and cow) of the three artiodactyl group together on the tree, whereas the dog and pig (the remaining artiodactyl) appear equally close to the cow-sheep pair (fig. 3). Indeed, the cow and sheep are identical at 97 of the 101 amino acid residues, while the dog and pig differ from the cow-sheep pair at 10 or 12 residues, respectively. Interestingly, IL8 is notably absent from the rat and mouse, although it is present in the guinea pig, another rodent. The guinea pig IL8 protein is the most divergent one in the group, having unique residues at 20 positions. Further, several of these guinea pig substitutions involve rather radical amino acid changes, such as K to Q (lysine to glutamine) at position 19, A to R (alanine to arginine) at position 21, and K to D (lysine to aspartic acid) at position 117.

The 16 GRO, MIP2, KC, and CINC2 sequences from eight species cluster together, and interspecific protein similarity ranges from 57% (between rat CINC2 and guinea pig KC) to 94% (between sheep GRO and cow GROy). The proteins in this group are most similar to one another in the 40-amino-acid region immediately following the ELR motif. The signal peptide and the final 30 amino acids near the C-terminus are more diverged between 66% and 87% amino acid similarity with each other and are more similar to each other in amino acid positions 80–120 in the C-terminus region of the protein than in the N-terminus. The PF4 genes reveal only 25%–48% similarity to any of the paralogous sequences. The PPBP proteins from humans and pigs share 67% similarity and are identical at 12 of 14 residues between positions 20–34 and at 24 of 28 residues between amino acids 92–119 (fig. 1). The IP10 proteins from humans, rats, and mice share about 70%–80% amino acid similarity. The human and mouse MIG proteins are 75% similar, while the human and cow GCP2 and the pig AMCFII share between 80%–90% amino acid similarity. The mouse and human MIG proteins share about 71% similarity and are longer than most chemokines, the complete coding regions in both species being about 125 amino acids. The two remaining proteins, the 114-amino-acid human ENA78 and the 132-amino-acid mouse LIX sequence, share about 60% amino acid similarity. The establishment of orthology among many of these comparisons is uncertain.

Relationships Among Paralogous Sequences

The maximum-likelihood tree suggests putative relationships among several groups of paralogous genes (fig. 3). It was of interest to test the statistical validity of these relationships by comparing the likelihoods of several alternative topologies to that of the maximum-likelihood topology (Kishino and Hasegawa 1989). As shown in table 2, five different sets of comparisons were carried out. The analysis presented in table 2A addresses the issue of whether or not the rodent KC and MIP2 genes are indeed orthologous to the GRO genes found in other species. This was done by comparing the likelihoods of subtrees in which these three proteins are depicted as each other’s closest relatives to subtrees in which these three proteins are shown to be more closely related to either PF4 or ENA78 than to each other. (Similar results were obtained when SDF1 or IP10 was used as the outgroup sequence). The two subtrees in which these three proteins are separated from each other both reveal statistically significantly ($P < 0.05$) lower likelihoods (difference in log likelihoods $= -11.1$ or $-11.2$) than do the subtrees that group these proteins together, indicating that KC and/or MIP2 are probably divergent GRO orthologs.

In table 2B–D, the validity of proposed sister taxon relationships between three pairs of sequences: PF4-PPBP, SDF1-IL8, and MIG-IP10, was examined. Al-
though separating any of these three pairs on the sub-
trees does result in a lower likelihood in each case, none of the lower likelihoods is statistically significant (al-
though the lower likelihood of 8.5 observed in sepa-
rating MIG and IP10 is nearly significant, and these two proteins do share two unique amino acid motifs not
found among other chemokines: EIIAT at positions 94–
98 and CLNPD/ES at positions 107–112. Similarly, PF4
and PPBP also share two unique amino acid motifs
that *K*_σ was about 30 substitutions per 100 synony-
mous sites among muroids, but it was up to 85 substitu-
tions per 100 sites in the rat–guinea pig comparison. *K*_σ
was about 9 substitutions per 100 sites among muroids,
but it was 32.5 per 100 sites in the mouse–guinea pig
comparison. Analyses failed to detect rate heteroge-
ity among the muroid KC genes (table 3). With respect
to the IL8 gene in primates and artiodactyls using
mouse comparisons) or guinea pig as an outgroup, and
analyses also failed to detect rate heterogeneity in the KC
genes among muroid taxa, discussed below), ENA78, and
GCP2.

Analyses of Rates of Nucleotide Substitutions

Relative-rate tests were conducted to test for pos-
sible rate heterogeneity in the KC genes among muroid
rodents (rats, mice, hamsters) using the hamster (in rat–
mouse comparisons) or guinea pig as an outgroup, and
among the IL8 genes in primates and artiodactyls using
various outgroups. Analyses of the KC genes revealed
that *K*_σ was about 30 substitutions per 100 synony-
mous sites among muroids, but it was up to 85 substitu-
tions per 100 sites in the rat–guinea pig comparison. *K*_σ
was about 9 substitutions per 100 sites among muroids,
but it was 32.5 per 100 sites in the mouse–guinea pig
comparison. Analyses failed to detect rate heteroge-
ity among the murid KC genes (table 3). With respect
to the IL8 gene in primates and artiodactyls, the *K*_σ
values ranged from 8.4 to 38.4 per 100 nonsynonymous
sites, while *K*_σ ranged from 3.2 to 15.1 per 100 sites. No
significant differences were found in any of the relative-
rate tests involving the IL8 comparisons (table 3).

The two monkey IL8 genes differ from one another
at 6 of 303 nucleotides, all of which cause amino acid
changes. As a result, between these two species \( K_s = \) 0.00 and \( \frac{K_a}{K_s} = 0.028 \pm 0.0172 \). In 11 other similar interspecific pairwise comparisons involving the IL8, GRO, KC, MIP2, IP10, and MIG genes, \( K_s \) was always significantly \((P < 0.05)\) larger than \( K_a \).

### Discussion

The genes in the \(\alpha\) chemokine subfamily share a number of features, including a genomic organization consisting of three or four exons (see Baggiolini, DeWald, and Moser 1994; Shirouzu et al. 1995); an open reading frame of about 100 amino acids (Matshushima et al. 1988); amino acid similarity ranging from about 25% to 80%; the presence of four conserved cysteine residues, with an intervening amino acid between the first two; cleavage of a signal peptide from the N-terminus prior to secretion from the cell (Van Damme et al. 1989; Yoshimura et al. 1989); the presence (in most cases) of a conserved ELR motif near the N-terminus (Baggiolini, DeWald, and Moser 1994); and a common chromosomal map position (all human genes have been mapped to chromosome 4q13–21 except the SDF1 genes, which are on chromosome 10) (Richmond et al. 1988; Shirouzu et al. 1995; Modi and Chen 1998). These findings suggest the origination of these genes from a common ancestral gene through repeated rounds of gene duplication.

The unequivocal establishment of orthology between divergent proteins (genes) from different species is not always straightforward. The present study analyzed primary amino acid sequences, and the results suggest that genes orthologous to many of the 11 known from humans have been isolated for other species. In addition, an assessment of the sister taxon relationships among paralogous genes was put forth, and it was concluded that many of the gene lineages were produced at approximately the same time. Further, it is useful to consider evidence derived from functional studies and chromosome mapping when evaluating the relationships among members of a large multigene family.

The analysis of IL8 sequences suggests that orthologous genes have now been isolated from nine species belonging to five different taxonomic orders. This establishment of orthology based on primary amino acid sequences is further supported by studies where that have shown that the IL8 protein from the rabbit chemoattracts neutrophils and increases the release of neutrophil elastase (Yoshimura and Yukhi 1991; Matsukawa et al. 1995), both featured characteristics of human IL8 activity. Curiously, however, this gene does not appear to be present in the genome of the rat or mouse, although it is present in the guinea pig (Yoshimura and Johnson 1993). Traditional taxonomists suggest that the rat-mouse and guinea pig (representing two different rodent suborders) lineages have been separated for about 55 Myr (Flynn, Jacobs, and Cheema 1986). What has apparently happened is that IL8 was lost in a common ancestor to the rat and mouse, and its function was adopted by another gene. It has been suggested that the mouse MIP2 or KC protein was the functional homolog of IL8 (Watanabe et al. 1993); however, from the protein sequence comparisons in figure 3, there is no evidence that MIP2 or KC and IL8 are orthologous genes. If future studies indeed reveal that these proteins carry out similar functions, then this would probably best be explained by convergent evolution or by the fact that these proteins may possess multiple functions.

The two Old World monkeys are thought to have last shared a common ancestor with one another about 10 MYA and with humans about 40 MYA (Benveniste 1985). Interestingly, the two monkey IL8 proteins are about as similar to one another as they are to the human sequence (about 95%), suggesting that selection has constrained this protein in primates. Further, the lack of synonymous differences \((K_s = 0.00)\) between the two monkey genes was unexpected. If we assume a neutral substitution rate of 0.1%–0.2% per Myr, then we would expect the two monkey genes to differ from one another by 2%–4%. Since the number of codons in the monkey IL8 genes is only 103, the lack of synonymous differences between them may be due to chance. In a comparison of 363 orthologous genes between rat and mouse, only one gene (IL-3) was found for which \(K_a\) was greater than \(K_s\) (Wolfe and Sharp 1993). Another example in which \(K_a\) exceeded \(K_s\) (\(d_a\) and \(d_s\) were actually calculated) was reported for the antigen recogni-

---

### Table 3

**Numbers of Synonymous \((K_s)\) and Nonsynonymous \((K_a)\) Nucelotide Substitutions per 100 Synonymous or Nonsynonymous Sites, Respectively, Among the KC and IL8 Genes**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Species 1</th>
<th>Species 2</th>
<th>Outgroup</th>
<th>(K_s)</th>
<th>(K_a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KC . . .</td>
<td>Rat</td>
<td>Mouse</td>
<td>Hamster</td>
<td>33.2 ± 11.4</td>
<td>9.3 ± 2.9</td>
</tr>
<tr>
<td>Rat</td>
<td>Mouse</td>
<td>Guinea pig</td>
<td>85.7 ± 25.7</td>
<td>30.6 ± 7.5</td>
<td>79.9 ± 21.5</td>
</tr>
<tr>
<td>Rat</td>
<td>Hamster</td>
<td>Guinea pig</td>
<td>85.7 ± 25.7</td>
<td>30.6 ± 7.5</td>
<td>75.8 ± 21.3</td>
</tr>
<tr>
<td>IL8 . . .</td>
<td>Mangabey</td>
<td>Macaca</td>
<td>Human</td>
<td>8.4 ± 4.7</td>
<td>3.9 ± 2.4</td>
</tr>
<tr>
<td>Human</td>
<td>Mangabey</td>
<td>Cow</td>
<td>38.4 ± 10.8</td>
<td>13.2 ± 4.3</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>Macaca</td>
<td>Rabbit</td>
<td>32.1 ± 9.7</td>
<td>11.8 ± 4.4</td>
<td></td>
</tr>
<tr>
<td>Sheep</td>
<td>Cow</td>
<td>Pig</td>
<td>27.9 ± 9.7</td>
<td>7.2 ± 2.5</td>
<td></td>
</tr>
<tr>
<td>Cow</td>
<td>Pig</td>
<td>Human</td>
<td>38.4 ± 10.8</td>
<td>13.2 ± 4.3</td>
<td></td>
</tr>
</tbody>
</table>

*Note:—Results of relative-rate tests (following Wu and Li 1985) in which each \(K_s\) and \(K_a\) value between the outgroup and species 1 were compared with each \(K_s\) and \(K_a\) value, respectively, between the outgroup and species 2 for a given gene were nonsignificant in all 16 comparisons. The number of codons examined for each gene ranged from 97 to 102.*
tion sites of the MHC class I molecules. This was interpreted to be attributable to overdominant selection, whereby increased protein polymorphism would be selectively advantageous in the heterozygote (Hughes and Nei 1988). The general lack of significant differences in $K_0$ observed in this study may denote a lack of selection or may be a statistical artifact, since a small number of codons were analyzed.

Multiple GRO genes have been identified for humans, cows, sheep, and rabbits, and several copies are probably present in other species as well. The three human sequences are very similar to each other, as are the three bovine proteins and two of the three rabbit proteins. Curiously, the third rabbit protein (GROγ) was observed to be more similar to the bovine sequences than to the two other rabbit proteins. These data (with the exception of the rabbit GROγ gene) would suggest that the duplication events producing the multiple genes within each of the different species either occurred convergently (and very recently, based on the extent of intraspecific similarity) in different taxa or preceded the diversification of these species, in which case the intraspecific similarity is attributable to gene conversion, recombination, or selection.

Among rodents, there are two clusters of proteins, KC and MIP2/CINC2, that appear to be the closest relatives of the GRO sequences found in the other species. The present analysis suggests that the rodent KC and/or MIP2 genes are divergent orthologs to human GRO and that CINC2 represents a paralogous gene that may currently carry out a different function. Additional evidence unifying GRO, KC, and MIP2 into a monophyletic clade comes from analyses of the transcriptional regulatory regions of these genes. For example, rat KC, mouse KC, mouse MIP2, and all three human GRO genes share a conserved κB1 motif that was shown to be essential for LPS inducibility and similar TATA boxes (Ohmori, Fukumoto, and Hamilton 1995). In addition, the mouse MIP2 gene shares SP1 and AP3 putative regulatory elements with the three human GRO genes (Widmer et al. 1993). Although the present analysis tentatively classifies the single guinea pig protein with the KC sequences from the other three rodents, it is not clear whether this guinea pig sequence is a divergent KC ortholog or actually a paralogous gene. Since other studies have also indicated unusually high rates of molecular evolution in the guinea pig (Li et al. 1992), additional sequences from other species of rodents and/or data regarding function are needed in order to better understand the relationships of the guinea pig protein. It is further curious that the rat and mouse both have KC and MIP2 genes, whereas our library screening efforts could identify only a single KC gene and no evidence of MIP2 in the guinea pig genome. In addition, the rat contains the CINC2 (α and β) genes. Whether or not orthologs of CINC2 are found in other species is currently unknown.

The fact that CINC2α and CINC2β differ from one another only by the presence of an additional three amino acids at the C-terminus of the β gene (Nakagawa et al. 1994) indicates that alternative splicing or the usage of different promoters is another mechanism for diversification in this gene subgroup. Alternative splicing has been shown to be responsible for the differences between the SDF1α and SDF1β genes in both the human and mouse genomes (Tashiro et al. 1993; Shirozu et al. 1995).

The divergent IP10 and MIG proteins, which exhibit minimal amino acid similarity to other members in the subfamily, cluster together as sister sequences in figure 3 just outside the level of statistical significance revealed by the Kishino-Hasagawa test (table 2D). However, independent evidence of kinship between these genes is available from other sources. Chromosome mapping both for humans (Modi and Chen 1998) and for mice (Modi et al. 1998) shows that these two genes are physically closer to each other than they are to any other genes in the α chemokine cluster. Further, both of these proteins utilize the same seven-transmembrane domain receptor, CXCR3, and are activated by interferon-gamma (Luster 1998). Along with PF4 and SDF1, MIG and IP10 lack the characteristic ELR amino acid motif located near the N-terminus of the other proteins that has been shown to be responsible for neutrophil chemotraction (Baggiolini, DeWald, and Moser 1994). These data suggest that the progenitor gene that gave rise to descendant chemokines via duplication lacked the ELR motif. However, the existence of a gene in chickens called 9E3 whose protein reveals about 30%–60% amino acid similarity to these mammalian chemokines that also contains the ELR motif, along with the conserved CXC and two additional C residues (Bedard et al. 1987), would argue that the common ancestor to the mammalian genes did possess the ELR motif. Therefore, the apparent loss of this motif has occurred in at least three convergent events during the radiation of this gene cluster if the topology in figure 3 is accurate.

The continuing discovery of novel CC chemokine genes (Van Coillie et al. 1997) suggests that additional CXC genes may also exist. Future analyses of new sequences may illustrate that the structural and evolutionary relationships among these sequences are even more complex.

Acknowledgments

Gary Smythers and Phil Poole, from the Frederick Supercomputer Center, provided assistance in running computer programs. Scott Davis and Noelle Cockett kindly provided genomic DNA samples. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

LITERATURE CITED


Phylogeny of Mammalian CXC Chemokines

VAN DAMME, J., J. VAN BEEUMEN, R. CONINGS, B. DECOCK, and A. BILLIAU. 1989. Purification of granulocyte chemo-
tactic peptide/interleukin-8 reveals N-terminal sequence heterogeneity similar to that of beta-thromboglobulin. Eur.
J. Biochem. 181:337–344.

WALZ, A. 1992. Generation and properties of neutrophil-acti-
vating peptide 2. Pp. 77–95 in M. BAGGOLINI and C. SORG,
eds. Interleukin 8 (NAP-1) and related chemotactic che-
mokines. Karger, Basel.

WALZ, Z., and M. BAGGOLINI. 1990. Generation of the neu-
 trophil activating peptide NAP-2 from platelet basic protein
from connective tissue-activating peptide III through mono-

WATANABE, K., M. IIDA, K. TAKAISHI, T. SUZUKI, Y. HAMADA,
Y. IIZUKA, and S. TSURUFUJI. 1993. Chemoattractants for
neutrophils in lipopolysaccharide-induced inflammatory ex-
udate from rats are not interleukin-8 counterparts but GRO-
gene-product/melanoma-growth-stimulating-activity-related

WIDMER, U., K. R. MANOGUE, A. CERAMI, and B. SHERRY.
1993. Genomic cloning and promoter analysis of macro-
phage inflammatory protein (MIP-2), MIP-1α, and MIP1β,
members of the chemokine superfamily of proinflammatory

WOLFE, K. H., and P. M. SHARPE. 1993. Mammalian gene evo-
lution: nucleotide sequence divergence between mouse and

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of

YOSHIMURA, T., and D. G. JOHNSON. 1993. cDNA cloning and
expression of guinea pig neutrophil attractant protein-1

YOSHIMURA, T., E. A. ROBINSON, E. APPELLA, K. MATUSHIMA,
Three forms of monocyte-derived neutrophil chemotactic
factor distinguished by different lengths of the amino-ter-

activation protein-1 and monocyte chemoattractant protein-
1 in rabbit: cDNA cloning and their expression in spleen

SIMPSON, reviewing editor

Accepted October 9, 1998