Discovery of Novel 4α helix Cytokine by Hidden Markov Model Analysis

Chunjuan Du[1,2,3] Yanjun Zeng[2] Yunping Zhu[3] Fuchu He[3]

1 School of Mathematical Sciences, South China Normal University, Guangzhou, 510631, P. R. China.

2 College of Life science and Bioengineering, Beijing University of Technology, Beijing, 100022, P. R. China.

3 Laboratory of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Taiping Road 27, Beijing, 100850, P. R. China.

E-mail: ducj_2001@yahoo.com.cn; zhuyp@hupo.org.cn; hefc@nic.bmi.ac.cn

ABSTRACT: Cytokines play a crucial role in the immune and inflammatory responses. But because of the high evolutionary rate of these proteins, the similarity between different members of their family is very low, which makes the identification of novel members of cytokines very difficult. According to this point, a new bioinformatic strategy to identify novel cytokine of the short-chain and long-chain 4α helix cytokine using hidden markov model (HMM) is proposed in the paper. As a result, two motifs were created on the two train data sets, which were used to search three different databases. In order to improve the result, a strict criterion is established to filter the novel cytokines in the subject proteins. Finally, according to their E-value, scores and the criterion, four subject proteins are predicted to be possible novel cytokines for each family respectively.

1 INTRODUCION

Since the identification of first cytokine, cytokine research has been always very hot in the field of biomedical research all over the world. Cytokines play a crucial role in the immune and inflammatory responses. So there would be a wide theoretical significance and practical application value to identify and study novel cytokines. Previously, cytokine research is almost completely dependent on experiment technique. Until the last years of 1990s, bioinformatical tools are much more involved in finding novel cytokines and giving some good advice. Taking interleukins (IL) for example, there were at least more than 10 novel interleukins identified in the past four years, such as IL19~IL32. The bioinformatic strategies, such as sequences analysis, database search and so on, play a key role in these researches. However, most of these methods depended on the proprietary and commercial EST databases, general researcher had little chance to explore it; those bioinformatic methods were only applied very simply, lacking of deep analysis. There still remains a lot of work to do for mining and extracting more valuable information from the cytokine related data.

The bottleneck of identifying novel cytokine is the low similarity (about 30%) of sequences because of high evolutionary rate. So the classical analysis tools for homology search such as BLAST [1] are not suitable for cytokine family. However, because of the similarity of function and structure, there would be some weak or long-distance similarity of cytokine sequences. Based on this point, a new bioinformatic strategy to identify novel cytokine of the short-chain 4α helix cytokine and long-chain 4α helix cytokine through constructing hidden markov model (HMM) is proposed in the paper. Two motifs are created on the two data sets of cytokine family, and which are used to search three databases (SwissProt, IPI, Nr,) for identifying those novel proteins related to the family. Then a criterion is established to filter the novel cytokines in the subject proteins. At last, four subject proteins are suggested to be novel cytokines for each family respectively combining E-value, scores and the criterion.

2 DATA AND METHODS

Two families of short-chain 4α helix cytokines and long-chain 4α helix cytokines are selected by structure classification according to SCOP [2] and related references. Both of them contain 14 members respectively (table 1), and are taken as train sets in the following studies. The proteins of the two train sets are from human or mouse, so each train set contains 28 sequences. Three famous protein sequences databases (SWISS-PROT, IPI and NR of NCBI) are used to search by each motif for identifying novel cytokine or cytokine-like protein.

The multiple alignments of the protein sequences were carried out using CLUSTALW [3] or TCOFFEE [4] and were edited manually for each tain set. Then, these aligned sequences were put into the hmmbuild program of the HMMER 2.3.2 package [5], and two HMM motifs were constructed by the two train sets. Following, hmmcalibrate program was used to calibrate the E-value and scores. The three databases are searched by each HMM motif using the hmmsearch program. For the probable novel cytokine, secondary structure is predicted by PSIPRED [6] and the known domain is analyzed by SMART [7]. In order to acquire the similarity between the predicted novel cytokine and the know cytokine of the two families, sequence alignments were performed by BLAST. For every protein including known cytokines and the subject proteins, the molecular weight, isoelectric point and hydrophobicity are computed by my own Perl program.

	members	number
short-chain 4α helix cytokines	IL2 IL4 IL5 GM-CSF M-CSF EPO TPO FLT3LG SCF IL3 IL13 IL15 IL21 IL31	14
long-chain 4α helix cytokines	GH G-CSF LIF CNTF PRL Leptin ONCM IL6 IL11 IL12p35 IL23p19 NNT1/CLC CT-1 IL27p28	14

Table 1: Short-chain 4α Helix and Long-chain 4α Helix Structure Cytokine Family

3 RESULTS AND DISCUSSION

3.1 The Database Search Results

The motif of 118aa was got for short-chain family and 266aa for long-chain family. The motif described the common characters of the cytokine family. After the three databases were searched by motifs, almost all proteins of train sets and all homologies of protein of train sets are contained in the results of database search for the two families (table 2 and table 3). For the short-chain family, all related cytokines (14 members) are contained in IPI and NR results. But in the SWISS-PROT results, there are five absent proteins including IL2, IL4, IL31 of human and IL15, IL31 of mouse. The absence of IL31 is because it is a new cytokine and not still embodied by SWISSPROT. For the long-chain family, every train protein has appeared in three database search results. At the same time, there are very few proteins that are not related to the short or long chain family. Most homologies of proteins of each train set are discovered by the search. Therefore, both the database search results of the two motifs have high sensitivity and specificity.

Short-Chain 4α helix Cytokines	SWISS-PROT (All species)	IPI (human)	NR (human)
Total	127	31	279
Protein of Train Set	23	14	14
Homology of Protein of Train Set	104	10	277
Probable Novel Protein	3	7	2
Non-redundancy		8	

Table 2: The Database Search Results of Short-Chain 4α helix Cytokine Famlily

After dealing with different database search results by removing redundancy proteins each other, 8 non-redundancy proteins are left for each family. For each set, based on cutoff of the default E value (10) and its score, 8 proteins are probable novel members. However, the setting of the default E- value and score depended on user's experience under common conditions. Because of the low similarity of cytokine, the proteins above the default E-value and score are not always the true member of the train set. In this instance, some strategy to filter real ones from the candidates is needed, which should reduce the false positive and improve the precision of prediction novel protein.

Long-Chain 4α helix Cytokines	SWISS-PROT (All species)	IPI (human)	NR (human)
Total	256	53	276
Protein of Train Set	14	14	14
Homology of Protein of Train Set	242	35	276
Probable Novel Proteins	0	18	0
Non-redundancy		8	

Table 3: The Database Search Results of Long-Chain 4α helix Cytokines

3.2 Criterion to predict novel cytokine

Usually, it is difficult to filter probable cytokine in the subject proteins without any known criterion. Here, we propose a criterion to solve the problem. On bases of E value and score, several cytokine characters are selected to filter the subject proteins such as secondary structure, chromosome location, sequence length, molecular weight, isoelectric point, hydrophobicity, known domain and so on.

The process has two steps. At first, E-value and score are the most important. E-value must be above the default cutoff and the best E-value is very low, which shows that the alignment between the motif and the subject protein in database is not by accident but homology. The best score is very high, which shows the alignment is very good and there are a lot of "identity or positive" positions. So, if both the E-value and score are better, the subject protein is more similar with the motif. Secondly, it is very crucial aspect whether biological characters of the subject proteins and the known cytokine are uniform. With more similar characters between subject proteins and the family, the match between subject protein and the motif is more credible, and the subject proteins are more possible to be novel cytokines. Cytokines are small secreted proteins, their molecular weights range from 15 to 30 kD. Their sequence similarities are low, but their secondary structures are very similar within the same cytokine family. Additionally, some genes coding cytokines always locate closely on the chromosome. And their molecular weight, isoelectric point and hydrophobicity are alike. Hence, a criterion including E-value, score, and biological features would ensure the precision and objectivity of prediction for cytokine-like protein. Moreover, the sequence similarity between the subject protein and the know cytokine of train set is very significant too. On my experience, if it is exceed 25% and the sequence length is not short than 40aa, the similarity is

believable for cytokine.

3.3 Eight novel proteins for two cytokine families

At last, 4 probable novel cytokines are filtered from the 8 non-redundancy subject proteins of each family using our criterion [Table 4, Table 5].

Protein	E-value and score	sizes aa	similarity
ATPF_BACME ATP synthase B chain	19.3 0.22	172	SCF (49aa,28%; 47aa, 27%)
Hypothetical Protein FLJ16535| (TMCC2_HUMAN)	0.1 3.3	709	IL2_MOUSE (75aa, 24%)
Stromelysin-1 precure (MMP3_RABIT)	3.2 5.9	478	IL31_MOUSE (47aa,38%) TPO (31aa, 35%)
Podocalyxin-like (PODX_HUMAN)	2.6 9.4	528	TPO (57aa, 33%)

Table 4: The four cytokine-like proteins of short-chain 4α helix cytokines

```
Score = 22.7 bits (47), Expect = 0.098
Identities = 14/49 (28%), Positives = 27/49 (54%)
s
ATPF_BACME: 29  LLALLQKFAFGPVMGIMKKREEHIAGE 55
              L +L+  FAFG +    KKR+  +
SCF_HUMAN: 223  LFSLIIGFAFGAL—YWKKRQPSLTRA 249

ATPF_BACME: 56  ID--EAEKQNEEAKKLVEEQRE 75
                ++  +  +++ E   L E++RE
SCF_HUMAN: 250  VENIQINEEDNEISMLQEKERE 269

Score = 21.6 bits (44), Expect = 0.22
Identities = 13/47 (27%), Positives = 25/47 (52%)

ATPF_BACME: 29  LLALLQKFAFGPVMGIMKKREEHIAGE 55
              L++L+  FAFG +    K+     A E
SCF_MOUSE: 223  LISLVIGFAFGALYWKKKQSSLTRAVE 249

ATPF_BACME: 56  IDEAEKQNEEAKKLVEEQRE 75
                  +  +++ E   L +++RE
SCF_MOUSE: 250  NIQINEEDNEISMLQQKERE 269
```

Figure 1 The Alignment of ATPF_BACME and SCF Using BLAST

The candidate proteins of short-chain 4α helix family almost all have α helix secondary structure except that there is little β sheet in MMP3_RABIT. The size of the 4 proteins all are within the scope of the known cytokines. From the E-value and score, ATPF is the most probable short 4α helix cytokine. There may be homologous between ATPF and SCF of human with 28% identity on 49aa sequence segment by BLAST and SCF of mouse with 27% identity on 47aa sequence sequent [Figure 1]. Because the similarity is very low among protein sequences of cytokine family, the BLAST result of ATPF and SCF shows that there may be remote homology between them and ATPF may be a novel short 4 α helix cytokine. The E-value is 0.1 and the score is 3.3 for TMCC2_HUMAN. Blast analysis shows TMCC2_HUMAN is just like IL2_MOUSE because of 24% similarity on 75aa sequence segment. For MMP3_RABIT and PODX_HUMAN, both the E-value and score are in the scope from 1 to 10. MMP3_RABIT is similar with IL31_mouse (47aa, 38%) and TPO (31aa, 35%), and PODX_HUMAN may have some relation to TPO (57aa, 33%).

Protein	E-value and score	sizes aa	similarity
26 KDA PROTEIN	-21.1 5	233	LIF (41aa, 29%; 32aa, 31%)
Amyloid beta A4 protein-binding family A member 2 binding protein (AB2BP)	-21.7 5.4	396	IL27_MOUSE (35aa, 56%)
Leucine zipper-ef-hand containing transmembrane protein 1	-22.0 5.7	739	ONCM (77aa, 31%; 75aa; 24%)
Hook homolog 2 (HOOK2)	-24.1 8	719	CTF1 (110aa, 30%; 66aa, 30%)

Table 5: The four cytokine-like proteins of long-chain 4α helix cytokines

```
Score = 22.7 bits (47), Expect = 0.27
Identities = 13/37 (35%), Positives = 21/37 (56%)

AB2BP: 210 GSSDTGRSSEAEMQWR 226
           G+  T  SSE E  W
IL27:  120 GTQGTWTSSEREQLWA 136

AB2BP:227 LQVNRLQELIDQLECKVRAVG 246
          ++++ L++L   L  +V A G
IL27: 137 MRLD-LRDLHRHLRFQVLAAG 155
```

Figure 2 The Alignment of AB2BP and IL27 Using BLAST

For long-chain 4α helix family, the secondary structure of the four candidate proteins all are α helix. The score-value are all lower than 0 and E-value are about 5 , so the alignment are not very reliable. But, through Blast analysis they are similar with some known members of long-chain 4α helix family. For example, AB2BP and IL27_MOUSE have 56% similarity on 35aa sequence

segment [Figure 2]. And HOOK2 is similar with CTF1 on 110aa about 30% [Figure 3]. It is long-distance similar between the Hook2 and CTF1_HUMAN, which is very infrequent unless they are homologous. So, both of them are believable to be homologous with cytokine. 26 KDA PROTEIN and LIF are alike and "Leucine zipper-ef-hand containing transmembrane protein1" is similar with ONCM. The other biological characters of four proteins is not opposite with the long-chain 4α helix cytokines. Therefore, the four protein are probable novel cytokines or cytokine-like proteins.

```
Score = 27.4 bits (59), Expect = 0.019
Identities = 34/110 (30%), Positives = 41/110 (36%), Gaps =
24/110 (21%)

Hook2:      400 VTKEKERLLAERDSLREANEELRCAQLQPRGL 432
                +TK  E+LL E   L           Q  P GL
CTF1_HUMAN:37  LTKYAEQLLQEYVQL----------QGDPFGL 69

Hook2:      433 TQADPSLDPTSTPVDNLAAEILPAELRE 459
                   PS  P   PV  L+A
CTF1_HUMAN: 70 ----PSFSPPRLPVAGLSAPAPSHAGLP 82

Hook2:      460 TLLRLQLENKRL----------CRQEAADRER 492
                  RL+L+   L           CR++A    R
CTF1_HUMAN:83  VHERLRLDAAALAALPPLLDAVCRRQAELNPR 115

Hook2:      493 QEELQRHLEDANRARHGL 499
                L R LEDA R    L
CTF1_HUMAN:116 APRLLRRLEDAARQARAL 132

Score = 23.9 bits (50), Expect = 0.22
Identities = 20/66 (30%), Positives = 26/66 (39%), Gaps = 13/66
(19%)

Hook2:      298 DELRQSSERAGQLEATLTSCRRRLGELRELRR 330
                + LRQ +     L A L + RRR  EL
CTF1_MOUSE: 85 ERLRQDAAALSVLPALLDAVRRRQAEL----- 112

Hook2:      331 QVRQLEERNAGHAERTRQLEDELRRAGSLRAQLE 363
                        N       R LED  R+  +L A +E
CTF1_MOUSE:113 --------NPRAPRLLRSLEDAARQVRALGAAVE 137
```

Figure 3 The Alignment Hook2 and CTF1 Using BLAST

3.4 Discussion

In a summary, the strategy of this paper has three parts. They are motif identification, database search and novel cytokine prediction. Each step is not limited to cytokines. Therefore, it is also applied on predicting novel member of any other protein family such as cytokine receptors [8] or any other low homologous protein families, if and only if the motif of family is efficient to search themselves and homologies of the datasets. Further experiment is necessary to validate the predicted function of these proteins, which will discovery and avoid false positive of prediction if it had occurred. severval novel cytokines are predicted and cytokine-like function of several ambiguous proteins is deduced objectively.

REFERENCES

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. Nucleic Acids Res. 25(17): 389--402, 1997

[2] A. Andreeva, D.Howorth, S.E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data. Nucl. Acid Res. 32:D226--D229, 2004

[3] D. Higgins, J. Thompson, T. Gibson J. D. Thompson et al. CLUSTALW: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673--4680, 1994

[4] O'Sullivan, K. Suhre, C. Abergel, D.G. Higgins, C. Notredame. 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. Journal of Molecular Biology. 340: 385--395, 2004

[5] Eddy S R. Profile hidden Markov models. Bioinformatics, 14:755--7, 1998

[6] L. J. McGuffin, K. Bryson, D. T. Jones. The PSIPRED protein structure prediction server. Bioinformatics. 16:404--405, 2000

[7] Ivica Letunic, Richard R. Copley, Steffen Schmidt, Francesca D. Ciccarelli, Tobias Doerks, Jörg Schultz2 Chris P. Ponting, Peer Bork. SMART 4.0: towards Genomic Data Integration. Nucleic Acids Research. 32: D142--D144, 2004

[8] Jerome A. Langer, E. Cali Cutrone a.1, Sergei Kotenko. The Class II cytokine receptor (CRF2) family: overview and patterns of receptor-ligand interactions. Cytokine & Growth Factor Reviews. 15, 33--48, 2004