



Protein Fingerprinting: A Domain-Free Approach to Protein Analysis

Deane fPa B, S. e l U e, Ca b da e, l^{1,*} Deane fM ec a B, M c b j a d B c en, S. e l U e, Ca b da e, l² Deane fPa, S a d A c a S j en, S. e l U e, Ca b da e, l³

¹Deane fPa B, S. e l U e, Ca b da e, l 62901, USA

²Deane fM ec a B, M c b j a d B c en, S. e l U e, Ca b da e, l 62901, USA

³Deane fPa, S a d A c a S j en, S. e l U e, Ca b da e, l 62901, USA

(—+—:2— 2004; — —:28. |—2004)

ABSTRACT: An alternative method for analyzing proteins is proposed. Currently, protein search engines available on the internet utilize domains (predefined sequences of amino acids) to align proteins. The method presented converts a protein sequence with the use of 1200 numeric codes that represent a unique three—amino-acid protein sequence. Each numeric code starts with one of three specific amino acids, followed by any two additional amino acids. With the use of the FPC (FingerPrinted Contig) program, the total protein database (including “redundant” records) from the National Center for Biotechnology Information (NCBI) has been processed and placed into “bins/contigs” based on associations of these triplet codes. When analyzed with FPC, proteins are “contigged” together based on the number of shared fragments, *regardless of order*. These associations were supported by additional analysis with the standard BLASTP utility from NCBI. Within the created contig sets, there are numerous examples of proteins (allotypes and orthotypes) that have evolved into different, seemingly unrelated proteins. The power of this domain-free technique has yet to be explored; however, the ability to bin proteins together with no *a priori* knowledge of domains may prove a powerful tool in the characterization of the hundreds of thousands of available, yet undescribed expressed protein and open reading frame sequences.

Keywords: Proteins, Fingerprinting, Domains, Contigs.

Table 2. Representation of amino acids in 1 million random protein records from NCBI.

Amino acid	Average per protein	Percentage of amino acids reported
W	4.317715	1.3027
C	6.464946	1.9505
H	7.81746	2.3586
M	8.189201	2.4708
Y	10.12765	3.0556
Q	12.54323	3.7844
F	13.51178	4.0766
N	15.19027	4.5831
D	16.69023	5.0356
P	17.10316	5.1602
R	17.57738	5.3033
K	17.94381	5.4138
T	18.52176	5.5882
I	18.92714	5.7105
E	20.26067	6.1129
V	20.78879	6.2722
G	22.3603	6.7463
S	23.31193	7.0335
A	28.69746	8.6583
L	31.09874	9.3828

In all, 331,443,618 amino acids were counted.

Table 3. Summary of nonredundant FASTA format protein processing, using the FPC program to "bin" triplet coded proteins together into contigs.

Number of proteins in dataset	800,171
Number of contigs	51535
Number of proteins in contigs	270,141
Average proteins per contig	5.24
Average number of triplets per protein	17.2

an illustration of highly related proteins from prokaryotes grouping together. Figure 3, Contig 29051, represents multiple bone morphogenesis proteins from several organisms and is an illustration of highly related proteins (including alternatively spliced isoforms) from eukaryotes contigging together. Descriptions of each protein were manually placed in the "Remarks" fields of clones from two sample contigs in the resulting FPC record.

The NCBI protein record gi5902813 was chosen for further investigation based on its location within FPC contig 29051 (Fig. 3) and is well described [27-35]. This record describes a protein-encoding locus that can induce cartilage formation and is reported to be identical to the

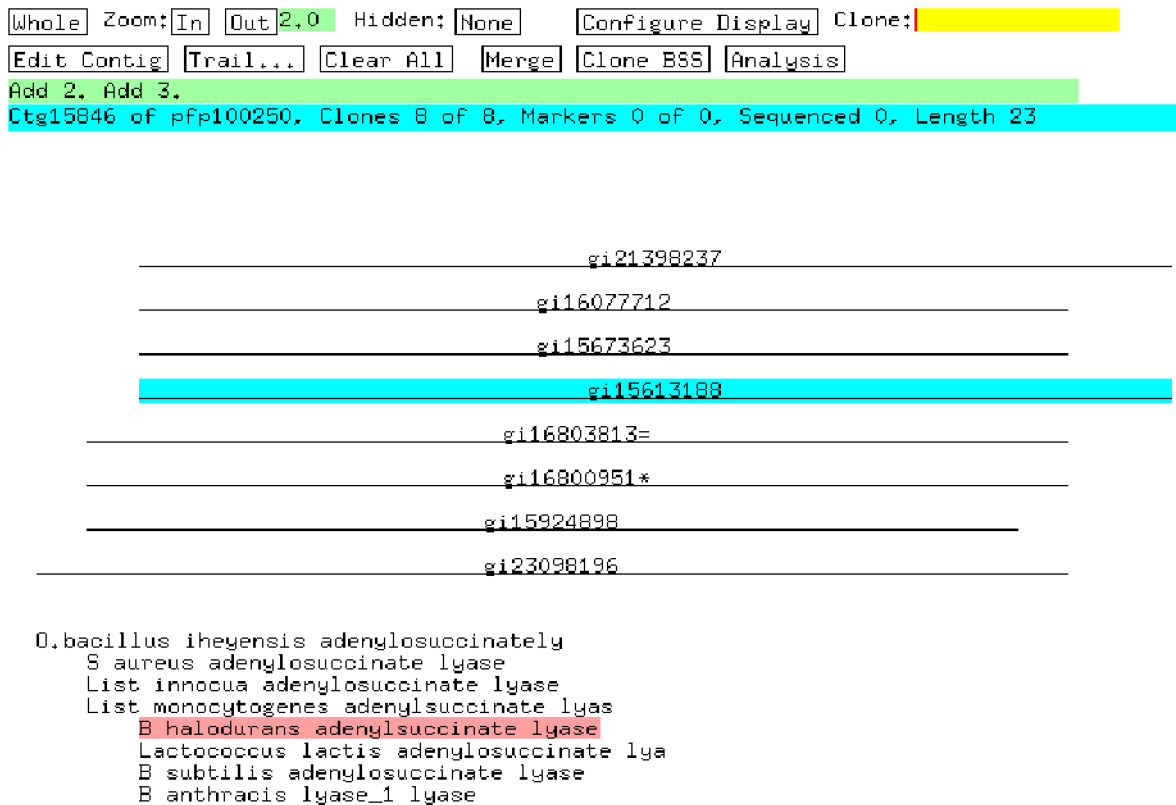


Figure 2. FPC illustration of contig 15846, showing seven similar adenylsuccinate lyase proteins and a similar protein from (gi 21398237, lyase 1). Data are shown in contiguous (a) and fingerprint (b) formats. Scale is in unique triplets.

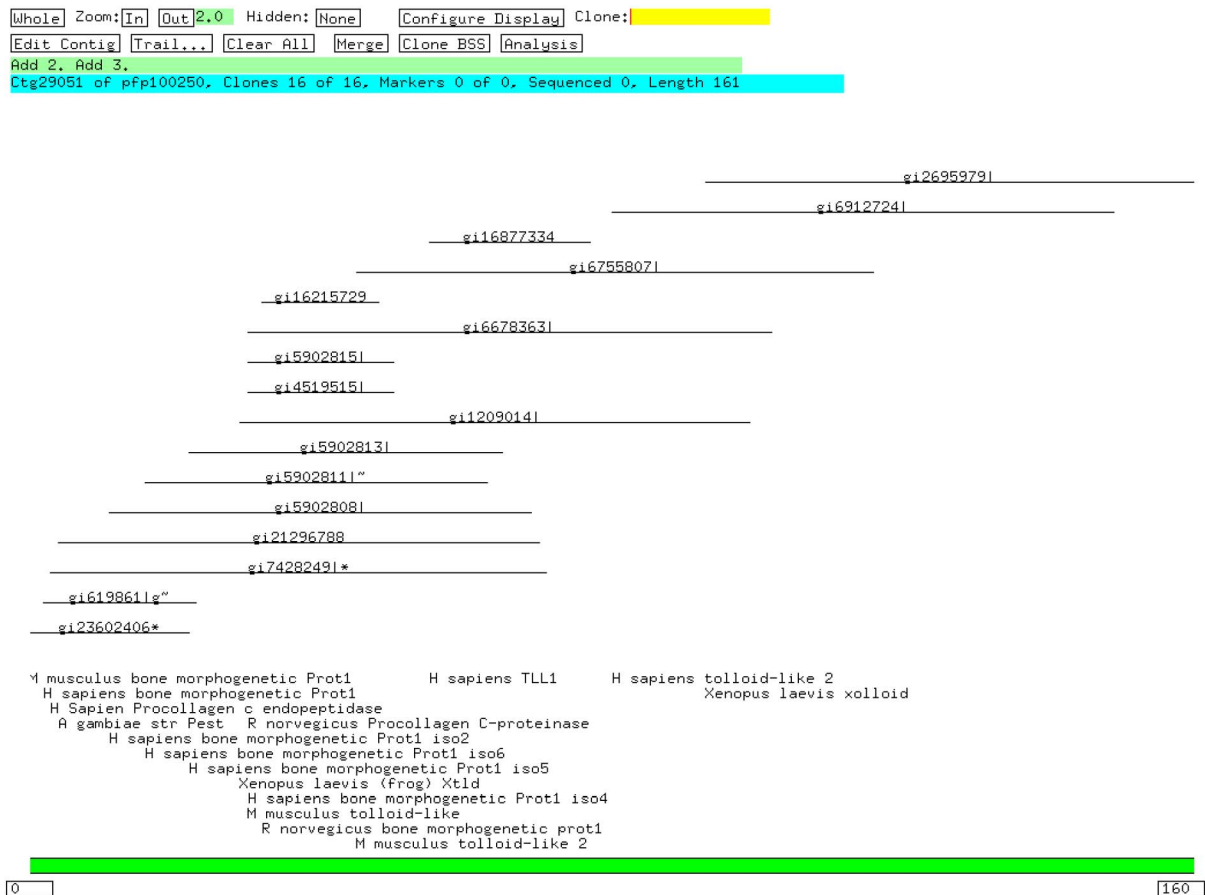


Figure 3. FPC illustration of contig 29051 showing several related proteins from human, rat, frog and mouse. This contig also includes an un-described protein from *Anopheles gambiae* str. gi21296788. Data is shown in contiguous (a) and fingerprint (b) formats. Scale is unique triplets.

secreted metalloprotease procollagen C proteinase (PCP). Expression of the BMP1 gene includes alternatively spliced variants that share N-terminal protease domains but may have varied C-terminal regions. An additional similarity investigation based on this record is shown in Figure 4.

As previously mentioned, protein similarity searching is also supported within the FPC program. By selecting the single-clone hitting tool, it is possible to search the entire dataset for matches, regardless of contig. The protein

gi5902813 from contig 29051 was used to test FPC, NCBI, and SwissProt search output. The top three hits from each search are listed in Table 4. It is clear that with the use of high cut-off parameters in FPC, the domain-free triplet code generates results equivalent to those available from NCBI and SwissProt.

A comparison of output after two iterations of the previously mentioned PSI-BLAST utility from NCBI using gi5902813 and the single clone hitting tool from FPC using

Table 4. Comparison of output from FPC, NCBI tBLASTn, and ExpASY BLAST for the random protein gi5902813 from the middle of contig 29051 (Fig. 3).

Database	Protein matches to gi5902813; bone morphogenetic protein 1 isoform 5, precursor; PCP [Homo sapiens]	(e)
FPC (tolerance 0, cutoff $5e^{-38}$)	GI:5902808 bone morphogenetic protein 1 isoform 2, precursor; PCP [, , ,] (contig 29051)	(0.0)
	GI:5902811 bone morphogenetic protein 1 isoform 6 precursor; PCP [, , ,] (contig 29051)	(0.0)
	GI:7428249 procollagen C-endopeptidase (EC 3.4.24.19) precursor, tolloid-like splice form-human (contig 29051)	(0.0)
NCBI	GI:1806029 . . . mRNA for bone morphogenetic protein BMP1-5	0.0
	GI:5902812 . . . bone morphogenetic protein 1 (BMP1), transcript variant BMP1-5, mRNA.	0.0
	GI:1806031 . . . mRNA for bone morphogenetic protein BMP1-6	0.0
ExpASY BLAST	P13497-4 Splice isoform BMP1-5 of P13497 Human	0.0
	P13497 BMP1_HUMAN Bone morphogenetic protein 1 precursor	0.0
	P13497-6 Splice isoform BMP1-7 of P13497 [BMP1] Human	0.0

The enclosed scores for FPC were based on the BLASTP 2 sequence comparison utility from NCBI when used to compare gi5902813 with the FPC hit.

the same record is shown in Figure 4a and b. Matches reported by FPC but not in the limited output from NCBI are compared with the use of NCBI BLASTp to obtain an e value for the pair, which is reported in Figure 4c.

4. CONCLUSIONS

The general frequency of each of the three initial amino acid residues (W, C, H) in different organisms [26] is supported by Table 2. Triplet WCM is the least represented in the dataset with 754 occurrences, and is closely followed by WWC (847) and CMW (867).

The triplet system allows major changes to occur without separating similar proteins because it measures less than 15% of the possible triplet combinations (1200 of 8000) within the proteins. The actual number is much lower than this; once frequency within proteins is accounted for, there is less than a 6% chance any random amino acid will be one of the three initial amino acids required to start encoding a triplet; $1.3\% (W) + 1.9\% (C) + 2.35\% (H) = 5.6\%$ (percentages from Table 2).

The processing time required to build (create contigs) is very high, usually 6–10 days. In addition, with a single cpu small steps in cutoff value are required to prevent crashing. These problems can be alleviated by applying greater computing power or designing new analysis software.

Using the single clone hitting utility in FPC requires no contig building step and yields data comparable to that of a BLAST search, with single clone searches providing accurate matches even at a cutoff of 1×10^{-10} . The two programs provide similar results at high e /cutoff values, but lower values from FPC do not correspond to those of NCBI, indicating either a loss in accuracy at low cutoff from FPC or incomplete similarities from NCBI.

Based on the example contigs and search comparisons, this method of protein analysis accurately portrays associations between proteins. The true power of this process lies in its flexibility and its ability to deal with changing primary structure.

The ability to compare, “bin,” and present proteins based on different parameters may help answer serious questions about whether (or when) an open reading frame is actually expressed by comparing with expressed proteins and therefore may help to elucidate protein relationships.

The analysis technique presented may help answer several questions: How many proteins are expressed by organisms? How many proteins are unique to each organism? Is the existence of similar proteins in different organisms evidence of necessity? Can phylogenetic relationships be determined on a gross scale with the use of varied protein sequence comparisons? The ability to “bin” proteins together with this flexible technique could lead to new insights into all of the above questions and forms the basis for further investigation.

Further work is continuing to create similar analysis procedures based on quadruplet codes. Full clone descriptions

and customized analysis tools for the Windows environment will also be incorporated. Although originally designed for presentation with the FPC program, a new interface designed to fully utilize advantages of this analysis technique is being developed.

The goal of this research is to provide investigators with a local, effective tool that allows modification of the analysis process itself to fit the unique structure of the protein being analyzed.

The JAVA code required to create the FPC file is available from the author.

The authors gratefully acknowledge Andrew Wood, Stephen Ebbs, David Gibson, Ahmad Fakhoury, Jorge Ferreira, and Khalid Meksem for their review of the manuscript and A. J. Afzal for his supportive conversations.

References and Notes

- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res*. 31, 23 (2003).
- O'Donovan C, Martin MJ, Glemet E, Codani JJ, Apweiler R. Removing redundancy in SWISS-PROT and TrEMBL. *Bioinformatics*. 15, 258 (1999).
- Apweiler R. 2001. Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. *Bioinformatics*. 2, 9 (2001).
- Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem Sci*. 23, 444 (1998).
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 215, 403 (1990).
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25, 3389 (1997).
- Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF. Protein sequence similarity searches using patterns as seeds. *Bioinformatics*. 26, 3986 (2010).
- Zhang H. Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm. *Bioinformatics*. 19, 1391 (2003).
- Baldi P, Chauvin Y, Hunkapiller T, McClure MA. Hidden Markov models of biological primary sequence information. *J Mol Biol*. 91, 1059 (1984).
- Brown M, Hughey R, Krogh A, Mian IS, Sjolander K, Haussler D. Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Bioinformatics*. 1, 47 (1986).
- Gribkov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *J Mol Biol*. 84, 4355 (1984).
- Hertz GZ, Hartzell GW, 3rd, Stormo GD. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Bioinformatics*. 6, 81 (1990).
- Loytynoja A, Miilinkovitch MC. A hidden Markov model for progressive multiple alignment. *Bioinformatics*. 19, 1505 (2003).
- MacKay Altman R. Assessing the goodness-of-fit of hidden Markov models. *Bioinformatics*. 60, 444 (2004).
- McLachlan AD. Analysis of gene duplication repeats in the myosin rod. *J Mol Biol*. 169, 15 (1983).
- Patthy L. Detecting homology of distantly related proteins with consensus sequences. *J Mol Biol*. 198, 567 (1987).

1. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *J. Mol. Biol.* 15, 1000 (1999).
1. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *J. Mol. Biol.* 12, 327 (1998).
1. Staden R. Computer methods to locate signals in nucleic acid sequences. *J. Mol. Biol.* 12, 505 (1974).
20. Tanaka H, Ishikawa M, Asai K, Konagaya A. 1993. Hidden Markov models and iterative aligners: study of their equivalence and possibilities. *J. Mol. Biol.* 1, 395 (1993).
21. Tatusov RL, Altschul SF, Koonin EV. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *J. Mol. Biol.* 91, 12091 (1994).
22. Taylor WR. Hierarchical method to align large numbers of biological sequences. *J. Mol. Biol.* 183, 456 (1985).
23. Yi TM, Lander ES. Recognition of related proteins by iterative template refinement (ITR). *J. Mol. Biol.* 3, 1315 (1994).
24. Soderlund C, Longden I, Mott R. FPC: a system for building contigs from restriction fingerprinted clones. *J. Mol. Biol.* 13, 523 (1994).
2. Soderlund C, Humphray S, Dunham A, French L. Contigs built with fingerprints, markers, and FPC V4.7. *J. Mol. Biol.* 10, 1772 (2000).
2. Takeuchi F, Futamura Y, Yoshikura H, Yamamoto K. Statistics of trinucleotides in coding sequences and evolution. *J. Mol. Biol.* 222, 139 (2003).
2. Garrigue-Antar L, Hartigan N, Kadler KE. Post-translational modification of bone morphogenetic protein-1 is required for secretion and stability of the protein. *J. Mol. Biol.* 277, 43327 (2002).
2. Hartigan N, Garrigue-Antar L, Kadler KE. Bone morphogenetic protein-1 (BMP-1). Identification of the minimal domain structure for procollagen C-proteinase activity. *J. Mol. Biol.* 278, 18045 (2003).
2. Janitz M, Heiser V, Bottcher U, Landt O, Lauster R. Three alternatively spliced variants of the gene coding for the human bone morphogenetic protein-1. *J. Mol. Biol.* 76, 141 (1993).
30. Kessler E, Takahara K, Biniaminov L, Brusel M, Greenspan DS. Bone morphogenetic protein-1: the type I procollagen C-proteinase. *J. Mol. Biol.* 271, 360 (1997).
31. Leighton M, Kadler KE. Paired basic/Furin-like proprotein convertase cleavage of Pro-BMP-1 in the trans-Golgi network. *J. Mol. Biol.* 278, 18478 (2003).
32. Li SW, Sieron AL, Fertala A, Hojima Y, Arnold WV, Prockop DJ. The C-proteinase that processes procollagens to fibrillar collagens is identical to the protein previously identified as bone morphogenetic protein-1. *J. Mol. Biol.* 93, 5127 (1993).
33. Tabas JA, Zasloff M, Wasmuth JJ, Emanuel BS, Altherr MR, McPherson JD, Wozney JM, Kaplan FS. Bone morphogenetic protein: chromosomal localization of human genes for BMP1, BMP2A, and BMP3. *J. Mol. Biol.* 9, 283 (1991).
34. Takahara K, Lyons GE, Greenspan DS. Bone morphogenetic protein-1 and a mammalian tolloid homologue (mTld) are encoded by alternatively spliced transcripts which are differentially expressed in some tissues. *J. Mol. Biol.* 269, 32572 (1994).
3. Takahara K, Lee S, Wood S, Greenspan DS. 1995. Structural organization and genetic localization of the human bone morphogenetic protein 1/mammalian tolloid gene. *J. Mol. Biol.* 29, 9 (1995).
3. Wozney JM, Rosen V, Celeste AJ, Mitscock LM, Whitters MJ, Kriz RW, Hewick RM, Wang EA. Novel regulators of bone formation: molecular clones and activities. *J. Mol. Biol.* 242, 1528 (1994).