2004

Protein Fingerprinting: A Domain-Free Approach to Protein Analysis

Jeffery L. Shultz
Southern Illinois University Carbondale

Chet Langin
Southern Illinois University Carbondale

Dennis G. Watson
Southern Illinois University Carbondale, dwatson@siu.edu

David Lightfoot
Southern Illinois University Carbondale, ga4082@siu.edu

Recommended Citation
Shultz, Jeffery L., Langin, Chet, Watson, Dennis G. and Lightfoot, David. "Protein Fingerprinting: A Domain-Free Approach to Protein Analysis." *Journal of Genome Science and Technology* 3, No. 1 (Jan 2004): 41-47. doi:10.1166/gl.2004.041.

Journal of Genome Science and Technology 2004, 3, 41–47
www.aspbs.com/gst

RESEARCH ARTICLE

Protein Fingerprinting: A Domain-Free Approach to Protein Analysis

Jeffry L. Shultz,[1,*] Chet Langin,[2] Dennis G. Watson,[3] and David Lightfoot[3]

[1] *Department of Plant Biology, Southern Illinois University, Carbondale, Illinois 62901, USA*
[2] *Department of Molecular Biology, Microbiology and Biochemistry, Southern Illinois University, Carbondale, Illinois 62901, USA*
[3] *Department of Plant, Soil and Agricultural Systems, Southern Illinois University, Carbondale, Illinois 62901, USA*

(Received: 2 March 2004; accepted: 28 June 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.

1. INTRODUCTION

The accelerating speed of protein discovery based on sequence analysis has created two problems. Millions of bits of genetic code are downloaded onto public-access databases every day [1], increasing the time required to search these databases. The number of perfect or near-perfect hits a user's query might yield [2] make the accuracy [3] and form of the information provided critical for further investigation.

Several protein analysis tools are available on the world wide web. The most common tool is the Basic Local Alignment Search Tool, or BLAST [4–6]. Performing a BLAST search, however, gives the stereotypical best hits based on a given sequence. Other tools, such as PHI-BLAST [7] and Longest Increasing Subsequence [8], have been presented to help solve this problem.

Distantly related proteins are now commonly grouped using Position-Specific Iterated BLAST (PSI-BLAST) [6], which utilizes well-described [9–23] position-specific score matrices (PSSM)/hidden Markov models (HMM).

The initial goal of protein fingerprinting was to graphically illustrate what is *not* the same from one similar protein to another. To facilitate this, proteins were disassembled into incomplete triplet amino acid sets (1200 of 8000 possible combinations) based on the starting amino acid being a tryptophan (W), cysteine (C), or histidine (H).

The analysis software most commonly available and designed to illustrate similarity in a simple format is FingerPrinted Contig, or FPC [24, 25], with version 6 being used to perform all calculations reported.

*Author to whom correspondence should be addressed.

 1551-7551/2004/01/041/007/$17.00+.25 doi:10.1166/gl.2004.041

A series of computational steps were performed to translate FASTA formatted proteins into FPC bands file format. Because it is incomplete and based on the relatively rare W, C, and H residues [26], this translation allowed great latitude for change within a protein to occur while still maintaining similarity.

A key advantage for this processing style is that the parameters can be changed by the user; for example, instead of W, C, and H, the user could specify I, F, and D, as the key amino acids. This ability to change the "key" amino acids and the ability of the FPC program to order and present proteins give the end user great flexibility. Essentially, each user can create his or her own domains and query the nonredundant protein database by using these new "domains."

2. MATERIALS AND METHODS

The entire nonredundant protein database from the National Center for Biotechnology Information (NCBI) was downloaded in Fasta format. Using Java, the 1.2-M nonredundant proteins file was processed to remove all identification but the "gi" identifying number and all interprotein text modifications (line feeds) and replace the selected triplets with a numeric code (Table 1). Each Fasta record was then sorted and written to a master FPC "bands" file. This process takes less than 10 min on a 2.0-GHz Pentium processor.

Samples of replacing protein sequence with fingerprint values are as follows:

FIAHFKLAFHKLHLRACSS FIA 3605 LAF 30522 32202 A 20909
LSCADKLCMLWSKGGFSLDF LS 2009 KL 2422 10905 GGFSLDF

Replacement of an amino acid is nonoverlapping and begins with the first occurrence of a W, C, or H and ends two amino acids later; thus the sequence "WWWRIT" yields the triplet "111 (*WWW*)" *not* "111 (*WWW*) + 1102 (*WWR*) + 10203 (*WRI*)." Any amino acids not included in a triplet are ignored by the conversion program.

A programming limit resulted in an error in 6 of 1200 coding sequences. This error occurs when CL or LC is in the second and third positions and produces the code number "#222." This sequence has occurred 45,011 times in this dataset (1222, 2222, or 3222 "band" size). Because this association will be correct in half of its occurrences, this is a 3 in 1200 chance or 0.25% error (the actual error rate is (45,011/2)/13,440,227 = 0.00167). This error will be corrected as a Windows-based interface is developed.

Identification of proteins with the "gi" reference number was necessary to account for limited name field size. Because of fixed-width data acquisition, some additional characters may appear after the protein ID, most commonly the "|" symbol.

With a dual 2.0-GHz processor computer with Linux 7.0 and version 6 of FPC, the processing time required to incrementally build or update the dataset of 800,000 proteins is usually 6–10 days. However, immediate queries of the data are possible without the creation of contigs and may more accurately reflect day-to-day usage of this dataset.

Table 1. Triplet codes used to replace protein sequence.

First		Second and third positions			
W	1	W	1	R	02
C	2	C	2	I	03
H	3	H	3	T	04
		M	4	K	05
		Y	5	V	06
		F	6	E	07
		N	7	G	08
		Q	8	S	09
		D	9	A	00
		P	01	L	22

As each unique FASTA record is processed, an occurrence of "W," "C," or "H" triggers the replacement of that residue with a 1, 2, or 3, then the following two amino acids are replaced according to the second and third position numeric values.

3. RESULTS AND DISCUSSION

A histogram of triplet frequency is shown in Figure 1, and a count of individual amino acids from 1 million random proteins in the initial dataset appears in Table 2.

As expected, the number of "bands" produced for each triplet increases with the overall frequency of the residue. The least frequent combination (750) was "124," corresponding to a "WCM" triplet; the most abundant combination was "32222," corresponding to "HLL."

A brief summary of the resulting data is presented in Table 3. The number of clones in contigs reflects a requirement for a multiple-step process in the integration of clones into contigs.

Examples of the contigs produced are shown in Figures 2 and 3. Contig 15846 (Fig. 2) represents multiple adenylosuccinate lyase proteins from several bacteria and is

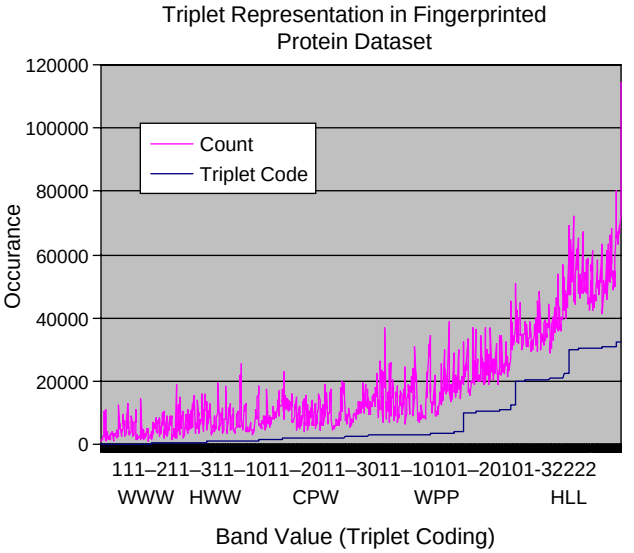

Figure 1. Triplet representation in a fingerprinted protein dataset.

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 *in vivo* and is reported to be identical to the

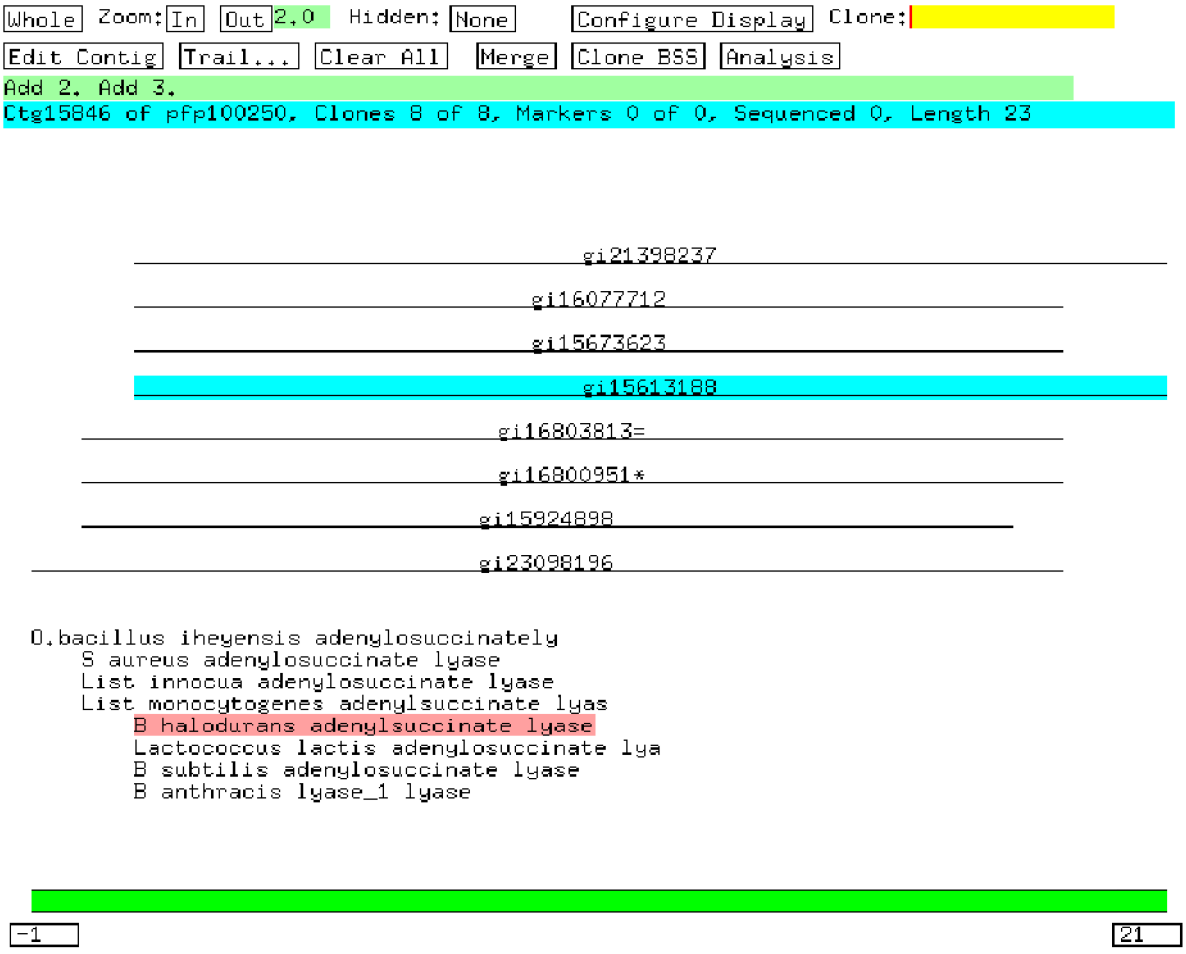

Figure 2. FPC illustration of contig 15486, showing seven similar adenylosuccinate lyase proteins and a similar protein from *B. anthracis* (gi 21398237, lyase 1). Data are shown in contiguous (a) and fingerprint (b) formats. Scale is in unique triplets.

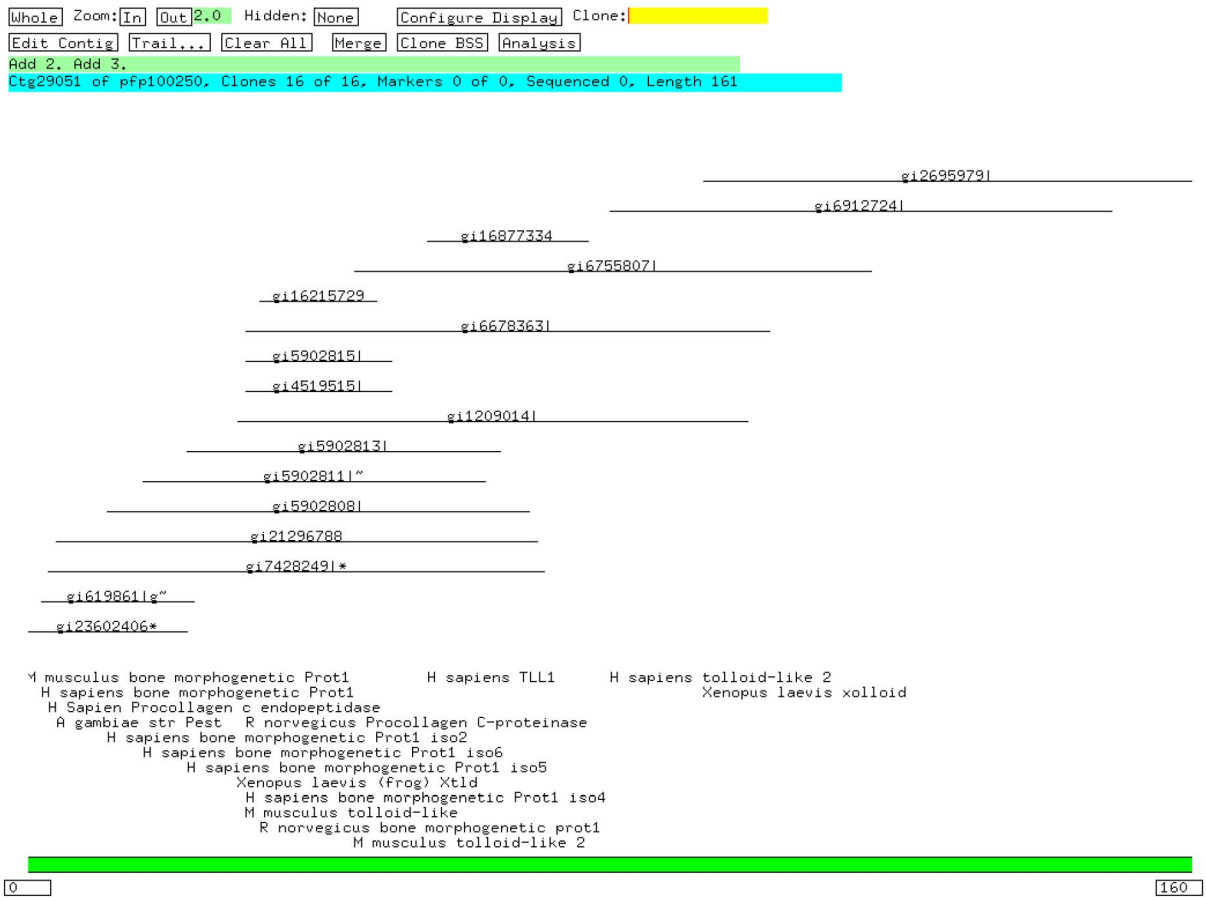

Figure 3. FPC illustration of contig 29051 showing several related proteins from human, rat, frog and mouse. This contig also includes an un-described protein from Anopheles gambiase str *Pest*, gi21296788. Data is shown in contiguous (a) and fingerprint (b) formats. Scale is unique triplets.

secreted metalloprotease procollogen 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 [*Homo sapiens*] (contig 29051)	(0.0)
	GI:5902811 bone morphogenetic protein 1 isoform 6 precursor; PCP [*Homo sapiens*].(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 *Homo sapiens* mRNA for bone morphogenetic protein BMP1-5	0.0
	GI:5902812 *Homo sapiens* bone morphogenetic protein 1 (BMP1), transcript variant BMP1-5, mRNA.	0.0
	GI:1806031 *Homo sapiens* 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.

(a)

```
>> gi5902813| ctg29051 44b --> Fpc  (Tol
0 Cutoff 1e-07)
Ctg29051      gi1209014| 71b    23 5e-27
Ctg0          gi1345615| 46b    14 7e-16
Ctg22535      gi1351101| 31b     8 6e-09
Ctg29051      gi2695979| 68b    21 4e-24
Ctg29051      gi4519515| 20b    10 3e-14
Ctg6876       gi4581057| 29b     7 9e-08
Ctg29051      gi5902808| 59b    39 5e-38
Ctg29051      gi5902811| 48b    39 5e-38
Ctg29051      gi5902815| 21b    17 7e-29
Ctg29051      gi6678363| 73b    22 4e-25
Ctg29051      gi6755807| 72b    22 3e-25
Ctg29051      gi6912724| 70b    23 3e-27
Ctg29051      gi7428249| 69b    39 7e-38
Ctg0          gi7511720| 74b    15 2e-14
Ctg0          gi15232881 29b     8 3e-09
Ctg0          gi15599363 28b     8 2e-09
Ctg29051      gi16215729 17b    15 9e-27
Ctg29051      gi16877334 22b    10 9e-14
Ctg0          gi16943691 16b     6 4e-08
Ctg0          gi18397646 17b     6 6e-08
Ctg20051      gi18846036 37b     8 3e-08
Ctg0          gi20094390 18b     6 9e-08
Ctg0          gi20544134 18b     6 9e-08
Ctg0          gi21290936 16b     7 7e-10
Ctg29051      gi21296788 66b    16 1e-16
Ctg49811      gi21450550 64b    13 1e-12
Ctg0          gi22988683 17b     6 6e-08
Ctg0          gi23051172 10b     5 1e-07
```

(b)

Proteins identified by FPC and not reported within the NCBI output in panel (a)

GI number	NCBI BLASTp Vs gi5902813	FPC cutoff value
1345615	0.0	7e-16
1351101	e-57	6e-09
4519515	e-57	3e-14
4581057	none	9e-08
15232881	none	3e-09
15599363	none	2e-09
16943691	e-06	4e-08
18397646	none	6e-08
18846036	none	3e-08
20094390	none	9e-08
20544134	6.6	9e-08
21290936	e-52	7e-10
21296788	e-164	1e-16
21450550	e-114	1e-12
22988683	none	6e-08

(c)

Figure 4. (a–b) Comparison of PSI-BLAST 2-iteration output and the output from FPC's single clone hit utility using accession gi5902813 from NCBI and Fingerprinted protein gi5902813| (identical record). The output form NCBI (a) is sorted by best match; the FPC output (b) is in sequential (random) order. BLASTP results form gi5902813 and clones listed in FPC output *not* matching NCBI output from (a) using the BLOSUM62 matrix are shown in (c).

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 e^{-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 *unique* proteins are expressed by *all* organisms? How many proteins are unique to each organism? Is the existence of similar proteins in different organisms *de facto* 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.

Acknowledgments: 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

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res.* 31, 23 **(2003)**.
2. O'Donovan C, Martin MJ, Glemet E, Codani JJ, Apweiler R. Removing redundancy in SWISS-PROT and TrEMBL. *Bioinformatics* 15, 258 **(1999)**.
3. Apweiler R. 2001. Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. *Brief Bioinform.* 2, 9 **(2001)**.
4. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* 23, 444 **(1998)**.
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 215, 403 **(1990)**.
6. 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)**.
7. Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* 26, 3986 **(1998)**.
8. Zhang H. Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm. *Bioinformatics* 19, 1391 **(2003)**.
9. Baldi P, Chauvin Y, Hunkapiller T, McClure MA. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA* 91, 1059 **(1994)**.
10. Brown M, Hughey R, Krogh A, Mian IS, Sjolander K, Haussler D. Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1, 47 **(1993)**.
11. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* 84, 4355 **(1987)**.
12. Hertz GZ, Hartzell GW, 3rd, Stormo GD. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* 6, 81 **(1990)**.
13. Loytynoja A, Milinkovitch MC. A hidden Markov model for progressive multiple alignment. *Bioinformatics* 19, 1505 **(2003)**.
14. MacKay Altman R. Assessing the goodness-of-fit of hidden Markov models. *Biometrics* 60, 444 **(2004)**.
15. McLachlan AD. Analysis of gene duplication repeats in the myosin rod. *J. Mol. Biol.* 169, 15 **(1983)**.
16. Patthy L. Detecting homology of distantly related proteins with consensus sequences. *J. Mol. Biol.* 198, 567 **(1987)**.

17. 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. *Bioinformatics* 15, 1000 **(1999)**.

18. 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. *Comput. Appl. Biosci.* 12, 327 **(1996)**.

19. Staden R. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* 12, 505 **(1984)**.

20. Tanaka H, Ishikawa M, Asai K, Konagaya A. 1993. Hidden Markov models and iterative aligners: study of their equivalence and possibilities. *Proc. Int. Conf. Intell. Syst. 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. *Proc. Natl. Acad. Sci. USA* 91, 12091 **(1994)**.

22. Taylor WR. Hierarchical method to align large numbers of biological sequences. *Methods Enzymol.* 183, 456 **(1990)**.

23. Yi TM, Lander ES. Recognition of related proteins by iterative template refinement (ITR). *Protein Sci.* 3, 1315 **(1994)**.

24. Soderlund C, Longden I, Mott R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* 13, 523 **(1997)**.

25. Soderlund C, Humphray S, Dunham A, French L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* 10, 1772 **(2000)**.

26. Takeuchi F, Futamura Y, Yoshikura H, Yamamoto K. Statistics of trinucleotides in coding sequences and evolution. *J. Theor. Biol.* 222, 139 **(2003)**.

27. 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. Biol. Chem.* 277, 43327 **(2002)**.

28. 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. Biol. Chem.* 278, 18045 **(2003)**.

29. 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. Med.* 76, 141 **(1998)**.

30. Kessler E, Takahara K, Biniaminov L, Brusel M, Greenspan DS. Bone morphogenetic protein-1: the type I procollagen C-proteinase. *Science* 271, 360 **(1996)**.

31. Leighton M, Kadler KE. Paired basic/Furin-like proprotein convertase cleavage of Pro-BMP-1 in the trans-Golgi network. *J. Biol. Chem.* 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 morphogenic protein-1. *Proc. Natl. Acad. Sci. USA* 93, 5127 **(1996)**.

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. *Genomics* 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. Biol. Chem.* 269, 32572 **(1994)**.

35. Takahara K, Lee S, Wood S, Greenspan DS. 1995. Structural organization and genetic localization of the human bone morphogenetic protein 1/mammalian tolloid gene. *Genomics* 29, 9 **(1995)**.

36. Wozney JM, Rosen V, Celeste AJ, Mitsock LM, Whitters MJ, Kriz RW, Hewick RM, Wang EA. Novel regulators of bone formation: molecular clones and activities. *Science* 242, 1528 **(1988)**.