Approaches to Detection of Distantly Related Proteins by Database Searches

BioTechniques 21:1118-1125 (December 1996)

K. Cattell, B. Koop, R.S. Olafson, M. Fellows, I. Bailey[1], R.W. Olafson and C. Upton
University of Victoria, and [1]BC Systems Corporation, Victoria, BC, Canada

ABSTRACT

The searching of protein databases as a method of identifying newly sequenced genes is commonplace in molecular biology laboratories. However, it is a procedure that is not usually formally taught to students, and method cookbooks discuss it only briefly. This article uses a single family of highly diverged uracil-DNA glycosylases, which fall into two distinct groups, to highlight some of the difficulties associated with identification of such proteins by database searching.

INTRODUCTION

The process of using a protein primary sequence to search protein databases for similar sequences is common to most molecular biology laboratories. The goal is usually to infer the function of a protein by its similarity to a previously characterized protein, thereby saving significant amounts of time and effort. Thus, database searching has become a very important tool in experimental protocols, and the process of database searching should be regarded as a complex experiment in itself.

There are numerous computer programs that are capable of searching protein databases for similar sequences, but results can differ dramatically. It is difficult and time-consuming to maintain multiple programs and current versions of the protein sequence databases in appropriate formats. Additionally, search programs are not always available in formats for all computers, or they may run slowly on older machines. These problems have been solved for many researchers with the introduction of database searching by remote computing. This involves sending data by e-mail to a powerful distant computer that is capable of performing the searches very rapidly, and which also uses e-mail to return the results. More recently, in some cases, it has become possible to access the remote computer by using the World Wide Web (WWW; Reference 3). Pedro Coutinho maintains an excellent catalogue of such sites, which can be accessed through the WWW (**http://www.public.iastate.edu/~pedro/research_tools.html**). Additional database searches may look for the presence of amino acid motifs or conserved domains, which are characteristic of many protein families. Such analyses may use the PROSITE (4) and BLOCKS (5) databases and rely on the previous characterization of a group of related proteins. Other approaches include the BEAUTY program (14), which attempts to integrate similarity and motif-based searching, and Blast3, which performs multiple alignments on sequences in the "hit list" (2).

In similarity searches, the time taken to search a database is dependent upon two invariant factors: the size of the database and size of the query protein; and upon two variables: the processing capacity of the computer and the particular program used. The use of "massively parallel" computer architectures addresses the first of these variables, although the choice of computer program and its associated parameter settings remains the most influential factor. Since the primary goal is the recognition of significantly related sequences that may have very low similarity, database searching protocols must attempt to optimize the speed of the search while providing maximum sensitivity.

Unfortunately, the output of database searches are not always simple to interpret. First, one should remember that there may not be any biologically significant similar proteins in the database, and that in such an example, the sequences at the top of the results list are there by chance. Second, also note that if two proteins are indeed related, several additional factors may still influence the success of database searching programs: *(i)* the overall percentage identity and percentage similarity between the proteins, *(ii)* the number of gaps required to align the regions of similarity and *(iii)* the distribution of the identical and similar residues within the protein sequences. As an example of the last point, it may be easier to recognize similarity between two proteins with 20% identity overall (global) if there is at least one small contiguous region

Table 1. Percentage Amino Acid Identity Among the Eight Uracil-DNA Glycosylases

	SFV	VV	FPV	HSV	EBV	Yeast	Strep.
VV	70						
FPV	59	55					
HSV	21	23	24				
EBV	17	18	20	42			
Yeast	18	21	21	44	40		
Strep.	22	21	24	34	39	41	
E. coli	19	21	23	45	47	51	47

(local, without gaps) with significantly higher similarity.

In this article, we wish to present some observations that significantly influence the success of detecting similarities between distantly related proteins and especially those without hot spots of high local similarity. These are in part based upon the identification of a poxviral uracil DNA glycosylase (12). While, with this example, we demonstrate the success of the programs NW_Align and FASTA, it is not our intention to rigorously compare the various search programs, but to use the results from searches with this protein family to illustrate a number of key points.

MATERIALS AND METHODS

Uracil DNA glycosylases from eight organisms were used in the database searches: Shope fibroma virus (SFV), vaccinia virus (VV), fowlpox virus (FPV), herpes simplex virus type-1 (HSV), Epstein Barr virus (EBV), *Saccharamyces cerevisiae* (yeast), *Streptococcus pneumoniae* (Strep.) and *Escherichia coli* (*E. coli*) with the following SWISS-PROT Database Accession numbers: P32941, P20536, P21968, P10186, P12888, P12887, P23379 and P12295, respectively. There are also a number of other uracil DNA glycosylases in the protein database.

Database searches (SWISS-PROT No. 32, containing 49 340 protein sequences) were performed at the following locations: *(i)* Blastp, **blast@ncbi.nlm.nih.gov** (BLOSUM62 matrix); *(ii)* Blitz, **Blitz@embl-heidelberg.de** (PAM120 matrix); *(iii)* FASTA v1.5, **fasta@genome.ad.jp** (PAM250 matrix; KTUP=1); *(iv)* DFlash, **dflash@watson.ibm.com** (PAM250 matrix); *(v)* FASTA v2.0 (BLOSUM50 matrix; KTUP=1) and *(vi)* NW_Align (RBO matrix) and were run locally on an Indy workstation (Silicon Graphics Inc., Mountain View, CA, USA). Help for the e-mail servers can be obtained by sending a message with "help" (no quotes) as the subject or as the message. Unless otherwise indicated, searches were performed using default parameters.

The NW_Align program is a module of the SEQSEE package (13), which runs on UNIX machines. It uses the Needleman-Wunsch algorithm (7), which scores global alignments and permits the introduction of gaps into the alignment. It also uses a nonstandard weight matrix (RBO), which was derived initially to function in similarity-based secondary structure prediction (13).

E-mail servers performed searches within highly variable times; often, results were not returned until the next day. FASTA searches, using v2.0 locally, took 3–4 min using KTUP=1. Blastp searches performed remotely on the WWW take less than 10 s of central processing unit (cpu) time, and results are usually returned within a minute, unless the load on the server is high (**http://www.ncbi.nlm.nih.gov/BLAST**). NW_Align searches, which were run with alignment scores sorted by raw score divided by sequence length, took approximately 1 h (e-mail: **seqsee@procyon.biochem.ualberta.ca** for further information).

RESULTS

It is instructive to review how this project began. Initial BLAST (1) and FASTA (9) database searches using an unidentified open reading frame (ORF) from SFV revealed similar ORFs in two other poxviruses: vaccinia virus and fowlpox virus. These three poxvirus proteins were clearly related, possessing amino acid identity ranging from 55% to 70% (Table 1), but no other significant matches were observed in the database searches.

Subsequent more rigorous searches using the newly available NW_Align module of SEQSEE as the search engine (13) placed several uracil-DNA glycosylases in the "hit list". While none of these uracil-DNA glycosylases produced the best score (after the poxvirus family of proteins), they were from very different organisms (HSV-1, yeast, *E. coli* and *S. pneumoniae*) and therefore unlikely to be highly similar. In fact, these enzymes share a low of 34% and a maximum of 51% identical amino acid residues (Table 1). In this situation, the significance of the matches was increased because the query detected homologous proteins that were themselves distantly related. Alignment of the poxvirus proteins with the uracil-DNA glycosylases (Figure 1) confirmed the matches detected by the NW_Align program were correct. Thus, two elements were critical in recognizing the match between the two sets of proteins: first, the search matched several members of a protein family, and second, the person running the search understood the significance of a single protein matching several members of protein family that were themselves distantly related.

The similarity between the original family of uracil-DNA glycosylases is limited to discrete regions of the protein sequences, and only a subset of these are also conserved in the poxviral proteins. Note that a number of small gaps must be inserted into the two groups of sequences to produce an optimal alignment (Figure 1). Subsequently, expressed protein from the SFV ORF demonstrated that this protein does possess the predicted uracil-DNA glycosylase activity (12). This represented the first DNA repair enzyme to be characterized in poxviruses and is especially interesting since it is essential for the replication of at least one poxvirus (10), whereas this enzyme is nonessential in other organisms. The PROSITE motif for uracil DNA glycosylases was subsequently modified to permit identification of the poxvirus enzymes.

To discover why the link between these two groups of homologous proteins was only detected by the NW_Align program, we have performed database searches using three

poxviral (group A) and five other (group B) uracil-DNA glycosylases as the query sequences with several different search programs available through the Internet. In these searches, we have simply scored the presence or absence of matches anywhere in the hit list returned with default parameters. Although this scoring system includes matches that may not normally be considered significant, it mimics the process by which we originally identified the poxvirus uracil-DNA glycosylases. As expected, all programs successfully matched each query sequence with the other members of the query group. However, when scored for the ability to match a uracil-DNA glycosylase from one group with members of the other group, the programs performed differently (Table 2). It is also clear that all of the sequences in a group did not produce similar search results with a particular program. The best example of this can be seen with the FPV sequence, which consistently performed better than the other two poxvirus sequences (Table 2). The Blastp search with this query detected three other uracil-DNA glycosylases from the other group (HSV, Strep. and *E. coli*), which would have been sufficient to call attention to these matches even if they were not very high in the hit list. Equivalent searches with the other two poxvirus proteins (SFV and VV) failed to match any of the other group. This result can probably be ascribed to the slightly higher similarity of the FPV protein to the members of the other group (Table 1). The FPV protein shares 0%–4% greater identity with members of the other group than do the SFV or VV proteins, and high scoring amino acid substitutions may also influence the results. Similarly, searches with the other query group frequently matched only the FPV protein in the poxvirus group. Such matches were usually low in the hit list, surrounded by insignificant matches, and therefore this single match would not be sufficient to allow a researcher to recognize the existence of a relationship between the two

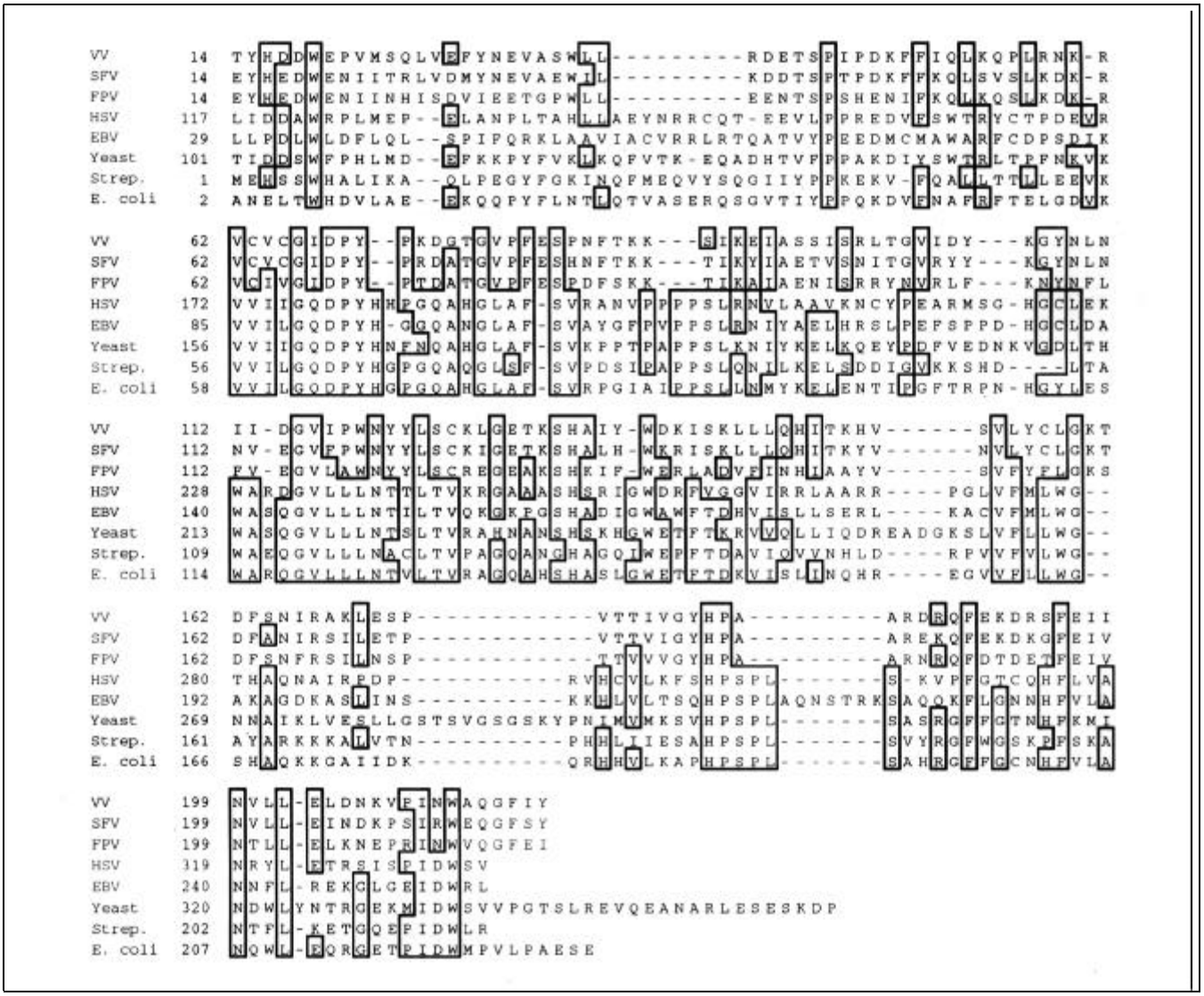

Figure 1. Alignment of the eight uracil-DNA glycosylases. The proteins have been truncated at the N-termini to match the start of the shortest sequence. Residues conserved in at least 50% of the sequences have been boxed. Protein sequences were manually aligned using SeqVu (Garvan Institute for Medical Research, Sydney, NSW, Australia).

Table 2. Detection of Uracil-DNA Glycosylases by Several Database Search Programs

	SFV	VV	FPV	HSV	EBV	Yeast	Strep.	*E. coli*
Blastp			HSV Strep. *E. coli*				FPV	FPV
FASTA v1.6	HSV	HSV	HSV *E. coli*	FPV VV				FPV
Blitz			HSV Strep. *E. coli*	FPV			FPV	FPV
DFlash			Strep.				FPV	
NW_Align	HSV Strep. *E. coli* Yeast	HSV Strep. *E. coli*	HSV Strep. *E. coli* Yeast	FPV VV SFV	FPV	FPV	FPV VV	FPV VV SFV

Each protein detected members within the same group. Therefore, only hits from the non-query group are reported. Default parameters were used.

Table 3. Detection of Uracil-DNA Glycosylases by BLITZ with a PAM240 Matrix

	SFV	VV	FPV	HSV	EBV	Yeast	Strep.	*E. coli*
Blitz	HSV *E. coli*	HSV *E. coli*	HSV *E. coli*	FPV VV			FPV VV	FPV VV
PAM240	Strep.	Strep.	Strep.	SFV			SFV	SFV

With all searches, each protein detected members within the same group. Therefore, only hits from the non-query group are reported.

protein groups. For this, it is necessary to produce hits with each of the poxvirus proteins and also be aware that these viruses are quite dissimilar. In this series of searches, the NW_Align program performed the best, since each of the poxvirus proteins were detected with the HSV and *E. coli* queries. NW_Align performs global alignments using the Needleman-Wunsch algorithm (7) and is well suited to this particular search situation. However, we have observed that Blastp and FASTA perform better than NW_Align when searching with short sequences (data not shown); therefore, one must tailor the search program to the task at hand.

Although it was not our intention to make an exhaustive comparison of the search programs as performed by others (8), we did change from the default comparison matrix in several instances. This had little effect on the search results with the notable exception of BLITZ, where changing to a PAM240 matrix significantly increased the sensitivity of the searches with both groups as the query sequences (Table 3). This was not unexpected since high-numbered PAM matrices are usually better for remotely related sequences, but this point emphasizes the need for users to be aware of a program's default settings. However, for most programs there was little to be gained by switching from the default matrices when searching for distantly related proteins.

The e-mail server used to perform FASTA searches used v1.5 of the program (currently runs v2.0); therefore, in an attempt to provide a current evaluation of FASTA, we repeated the searches using FASTA v2.0 on a local machine. This version of FASTA uses the BLOSUM50 matrix and includes statistical analysis of the search. The results shown in Table 4 clearly demonstrate an improvement in the detection of these distantly related proteins. However, again this shows the importance of which protein is used as the query sequence. For example, searching with the yeast protein only detected one of the poxvirus proteins, whereas the others in the group detected two or all of the poxvirus proteins. A further difference between the query sequences can be observed by examining where the hits are ranked in the hit list (after removal of hits from search query group of proteins). Searching with the *E. coli* or HSV protein resulted in the detection of all three poxvirus uracil-DNA glycosylases (Table 4). However, the *E. coli* query placed the proteins 1, 2 and 3 in the hit list, while HSV placed them 1, 9 and 20. This is important since the actual alignments are weak, and the significance of the results must be analyzed by the researcher scanning the hit list for related sequences. It is therefore essential that the search program returns a sufficient number of hits to allow this visual scanning. One could envision that a chance hit against a protein with many very close homologues in the database could push other significant hits so far down the hit list that they are not returned by the search program or are disregarded by the researcher.

Table 4. Detection of Uracil-DNA Glycosylases by FASTA v2.0

	SFV	VV	FPV	HSV	EBV	Yeast	Strep.	*E. coli*
FASTA	HSV	HSV	HSV	FPV	FPV	VV	FPV	FPV
	E. coli	*E. coli*	*E. coli*	VV	VV		VV	VV
v2.0	EBV	EBV	EBV	SFV			SFV	SFV
	Strep.	Strep.	Strep.					
		Yeast						
Position	1	1	1	1	3	14	1	1
in	2	2	2	9	26		2	2
Hit list	3	3	3	20			3	3
	5	5	4					
		12						

With all searches, each protein detected members within the same group. Therefore, only hits from the non-query group are reported. Position in the hit list is given after the removal of members of the query group.

DISCUSSION

The identification of distantly related protein sequences is important from many viewpoints. We have used an extreme example to provide insights into how best one can go about detecting such relationships. However, it will always be a difficult and frustrating task because it is hard to know when a negative result is a true negative, and one cannot afford to waste time chasing false positives. Uracil-DNA glycosylases are present in very diverse organisms and have probably evolved from a single ancestral gene. While alignment of these proteins demonstrates that there are a number of highly conserved amino acids, they are distributed throughout the protein sequence, and there are no large blocks that are conserved amongst all members of the family. A further complication results from the need to introduce a number of short gaps into the groups of sequences for optimal alignment (Figure 1). Indeed, these gaps are most likely responsible for the relatively poor results using Blastp with this particular family of proteins. A crystallographic study of one of the enzymes demonstrated that the active site is formed from noncontiguous regions (6), and this may account for the lack of any highly conserved block. This family of proteins may prove to be a useful control group in other tests of database search programs.

Our example of the identification of the poxvirus uracil-DNA glycosylases is not a unique situation. A poxvirus interferon-gamma binding protein was identified in an identical manner (11). These poxviral proteins are believed to have evolved from a eukaryotic interferon-gamma receptor, but the similarity is very low. Currently, Blastp does not detect these proteins, but NW_Align and FASTA match the viral proteins with both the human and mouse interferon-gamma receptors. Again, the significance of these matches is enhanced by an appreciation of the relatively low similarity between the human and mouse receptors.

It is not our intention to suggest that any one search program is always better than another. While the various algorithms, matrices and implementations clearly produce different results, our aim was to highlight the less obvious differences and suggest approaches to database searching that might increase the chances of detecting distantly related sequences. These are equally pertinent when searching the databases with a single unknown protein or with a group of proteins looking for new family members.

Table 5. Summary of Observations

1) The person performing the search can provide valuable insight into the biological significance of sequences in the hit list.
2) The significant matches are not always at the top of the hit list.
3) If working with a group of unidentified proteins, then searches should be performed with all members of the group. Not all proteins within a family may be able to match with all other members of the family in a standard database search.
4) When using a group of known proteins to look for new members of a family, the new candidate members themselves should be used to search the database since they are likely to match several members of the family if they are true members.
5) Searches should return at least the top 50 matches to permit manual scanning for similar proteins that are not closely grouped. Large families of unrelated proteins may force weak matches so far down the hit list that it appears unlikely that a protein with such a low ranking could be significantly related to the query.
6) Searches should be repeated using different programs.
7) Searches should be repeated at regular intervals because of the rapid growth of the databases. By the time one reads “no matches were observed in the database” in a journal, this result is out of date.
8) Current versions of the search programs and databases should be used. This information is not always obvious from the search results or help files.
9) Beware of false positives.

Our observations are summarized in Table 5.

In conclusion, since large amounts of resources are used in the determination of DNA/protein sequences, one should be prepared to expend a significant amount of time on the analysis of this data.

ACKNOWLEDGMENTS

This work was funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Operating Grant to C.U. and Canadian Genome Analysis and Technology and NSERC grants to B.K. The help of the BC Systems Corporation is gratefully acknowledged. We would like to thank David Wishart and Robert Boyko for help with NW_Align.

REFERENCES

1. **Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. *215*:403-410.
2. **Altschul, S.F. and D.J. Lipman.** 1990. Protein database searches for multiple alignments. Proc. Natl. Acad. Sci. USA *87*:5509-5513.
3. **Atwell, R., F. Gibbins and C. Upton.** 1995. Using a World Wide Web server as a local organizer for protein and DNA sequences. BioTechniques *19*:966-970.
4. **Bairoch, A.** 1991. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res. *19*:2241-2245.
5. **Henikoff, S. and J.G. Henikoff.** 1994. Protein family classification based on searching a database of blocks. Genomics *19*:97-107.
6. **Mol, C.D., A.S. Arvai, G. Slupphaug, B. Kavli, I. Alseth, H.E. Krokan and J.A. Tainer.** 1995. Crystal structure and mutational analysis of human uracil-DNA glycosylase: structural basis for specificity and catalysis. Cell *80*:869-878.
7. **Needleman, S.B. and C.D. Wunsch.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*:443-453.
8. **Pearson, W.R.** 1995. Comparison of methods for searching protein sequence databases. Protein Sci. *4*:1145-1160.
9. **Pearson, W.R. and D.J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA *85*:2444-2448.
10. **Stuart, D.T., C. Upton, M.A. Hilman, E.G. Niles and G. McFadden.** 1993. A poxvirus-encoded uracil DNA glycosylase is essential for virus viability. J. Virol. *67*:2503-2512.
11. **Upton, C., K. Mossman and G. McFadden.** 1992. Encoding of a homolog of the IFN-gamma receptor by myxoma virus. Science *258*:1369-1373.
12. **Upton, C., D.T. Stuart and G. McFadden.** 1993. Identification of a poxvirus gene encoding a uracil DNA glycosylase. Proc. Natl. Acad. Sci. USA *90*:4518-4522.
13. **Wishart, D.S., R.F. Boyko, L. Willard, F.M. Richards and B.D. Sykes.** 1994. SEQSEE: a comprehensive program suite for protein sequence analysis. CABIOS *10*:121-132.
14. **Worley, K.C., B.A. Wiese and R.F. Smith.** 1995. BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. Genome Res. *5*:173-184.

Received 3 April 1996; accepted 26 July 1996.

Address correspondence to:

Chris Upton
Department of Biochemistry and Microbiology
Room 150 Petch Building
University of Victoria
Victoria, BC V8W 3P6, Canada
Internet: cupton@uvic.ca