Fall 2002

A method for developing in-silico protein homologs

Susan McClatchy
New Jersey Institute of Technology

ABSTRACT

A METHOD FOR DEVELOPING *IN-SILICO* PROTEIN HOMOLOGS

by
Susan McClatchy

Computational methods for identifying and screening the most promising drug receptor candidates in the human genome are of great interest to drug discovery researchers. Successful methods will accurately identify and narrow the field of potential drug receptor candidates. This study details one such method.

The method described here begins with the assumption that novel drug receptors have high sequence similarity to established drug receptors. The similarity search program FASTA3 aligns translated sequences of the human genome to known drug receptor sequences and ranks these alignments by measuring their statistical significance. Query results returned by FASTA3 are assembled into "*in-silico* proteins" or artificially generated homologs of known drug receptors. A second similarity search program, BLASTP, aligns *in-silico* proteins with a protein database, and also ranks alignments based on statistical significance. A potentially valuable *in-silico* protein identifies its generating drug receptor as the top-ranking result returned from the BLASTP search, and may represent a new family member of a particular group of drug receptors.

A METHOD FOR DEVELOPING *IN-SILICO* PROTEIN HOMOLOGS

by
Susan McClatchy

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey - Newark
In Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Federated Biological Sciences Department

January 2003

APPROVAL PAGE

A METHOD FOR DEVELOPING IN-SILICO PROTEIN HOMOLOGS

Susan McClatchy

Dr. Michael Recce, Thesis Advisor — Date
Associate Professor of Information Systems, NJIT

Dr. Barry Cohen — Date
Assistant Professor of Computer Science, NJIT

Dr. Alex Elbrecht — Date
Senior Research Fellow, Merck Research Labs, Rahway, NJ

BIOGRAPHICAL SKETCH

Author: Susan McClatchy

Degree: Master of Science

Date: January 2003

Undergraduate and Graduate Education:

- Master of Science in Computational Biology,
  New Jersey Institute of Technology, Newark, New Jersey, 2003

- Teacher Preparation Program,
  Humboldt State University, Arcata, California, 1991

- Bachelor of Science in Biology,
  Humboldt State University, Arcata, California, 1988

Major: Computational Biology

Dedicated to my parents, Al and Marion,
for all their love and support

No seas celoso sol

Me miró y preguntó el bendito sol,
¿Qué es lo tuyo triste alma,
es que no te conformas con el alba
que te da mi puro y eterno amor?

y le dije... "no seas celoso sol,
que tú todo lo tienes en el cielo,
mientras yo sólo pido el ardiente fuego
de un puro y sincero corazón."

~Jonathan Hernandez
1983 - 2002

ACKNOWLEDGMENT

My sincere thanks go to Dr. Michael Recce for his valuable advice and generous support throughout my studies at New Jersey Institute of Technology, and to Dr. Alex Elbrecht of Merck Research Labs for providing me with a clear direction and an unrestricted opportunity to learn.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Objective

The completion of a draft sequence of the human genome in June of 2000 [International Human Genome Sequencing Consortium, 2001] amplified efforts aimed at detection of genes serving as receptors for drugs. These efforts have produced a multitude of promising new candidate drug receptors that must be filtered thoroughly to find those best suited for new drug development. Ultimately, those genes whose products make the best leads for drug development must undergo rigorous experimental screening and analysis. Prior to experimentation, computational methods may be employed to screen and identify those genes most likely to serve as leads. This thesis details one computational method for identifying and screening novel drug receptor genes.

The project described here aims to define and implement a procedure for finding new members of any protein class, although it will focus on drug receptors. This procedure creates *in-silico* proteins, or artificial homologs, from query results obtained by searching a translation of the human genome with known drug receptor sequences, using the similarity search program FASTA3. An *in-silico* protein is verified if the BLASTP similarity search program finds that it has highest similarity to the known drug receptor from which it was derived.

Potential new members of a particular family of drug receptors may be found among these verified *in-silico* proteins.

1.2 Background Information

A concise definition of a drug receptor begins with the drug receptor model [Hardman, et. al., 2002]. According to this model, a drug binds to a specific receptor to form a drug-receptor complex, leading to some alteration of physiological function. A drug with therapeutic value causes some desired alteration of function, while undesired changes remain at an acceptable minimum. In pharmaceutical parlance the phrase "drug target" is commonly used in place of the phrase "drug receptor". The two phrases will be used interchangeably from here on.

Most of the targets of marketed drugs are proteins whose role in biological pathways may or may not be known. A focused search for new genomic drug targets may include a hunt for relatives of genes involved in disease pathways, or relatives of genes whose products are known to bind drugs therapeutically. Computational searches for new drug targets frequently rely on methods for identifying relatives by gene or protein sequence homology or protein structural similarity.

These computational searches lie at the beginning of a continuum that includes target screening, identification and validation [Branca, 2001 and

Swindells & Overington, 2002]. Methods such as gene sequence and gene or protein expression data mining provide minimal confidence, while experimental methods, such as *in vivo* functional studies of genes inserted the mouse genome (knockout mice), provide maximal confidence in the validation of a target. Computational methods offer correlation between a supposed target and its role in disease or therapy, while experimental methods imply direct causation [Federsel, 2001]. The value of computational methods lie in their capacity to screen large numbers of genes at once, eliminating many as poor candidates and decreasing the labor required further downstream, where targets must be validated in laboratory studies.

Correlative methods include use of gene expression microarrays to identify genes with differential messenger RNA (mRNA) expression in various states. A simple example of expression analysis would note differences in gene expression between diseased and healthy tissue, or between treated and untreated tissues. Serial Analysis of Gene Expression (SAGE) identifies genes and quantifies gene expression by joining several short segments, or tags, of mRNA from different genes. The joined tags are sequenced, identified, and counted to create an expression profile for genes known and unknown.

Mining of expressed sequence tag (EST) databases may be used to identify novel genes. Expressed sequence tags are partial sequences of complementary DNAs produced from reverse transcription of messenger RNA, and represent genes expressed in a particular tissue under certain conditions. Expressed

sequence tags may also be used to detect single nucleotide polymorphisms (SNPs) in disease genes. Single nucleotide polymorphism association studies compare disease and control populations, and compare allele frequency or genetic variation between two groups.

Proteomics methods involve proteins rather than genes, and include techniques for structural homology and mining of protein expression data. Using the assumption that sequence similarity denotes structural similarity, homology models may be generated for an unknown protein with high sequence similarity to a well-characterized protein. For low sequence similarity, threading methods are used. These structure-based methods may be useful in identifying likely binding sites for drugs.

Protein expression data report on differential expression of a gene's final product rather than the intermediary mRNA. Differential protein expression, like differential gene expression, may be used to evaluate tissues in a disease state versus a normal state. However, difficulties in purifying proteins hamper this more direct means of measuring expression levels.

Most searches for new drug targets utilize sequence similarity search programs. Resources such as HMMER and Pfam employ profile hidden Markov models (HMMs) of protein domains or conserved regions to find new protein family members [Bateman, et. al., 2002]. Hidden Markov models, originally applied to speech recognition, define the probability of a given sequence of states

and symbols [Durbin, et. al., 2001]. An HMM may describe, for example, the probability that a given sequence of amino acids (symbols) forms either an alpha helix or a beta pleated sheet (state). Profile HMMs best describe multiple alignments of protein family members, and are better suited to finding new members fitting a particular profile than are pairwise alignment tools such as FASTA and BLAST.

Programs like BLAST and FASTA measure the statistical significance of a local alignment between two sequences, rather than simply measuring percent sequence identity. Evaluating an alignment by percent sequence identity alone fails to recognize that short sequences may have very high sequence identity simply by chance [Wood & Pearson, 1999]. These latter two programs were used extensively to create *in-silico* proteins, and are described in further detail.

1.2.1 FASTA

In aligning sequences, FASTA creates a lookup table to rapidly locate identities between two sequences [Lipman & Pearson, 1985]. The name and position of each residue is maintained in the table, and offset values calculated for each pair of identical residues. An example follows.

	Position					
Sequence	1	2	3	4	5	6
1	A	T	A	R	G	A
2	R	C	R	G	A	Q

Example 1.1 Position offsets for identical residues are calculated and stored in a lookup table.

The lookup table indicates that for Sequence 1, A is found at Positions 1, 3 and 6, R at Position 4 and G at 5. Positions of residues in Sequence 2 are compared to those in Sequence 1. So, the R at Positions 1 and 3 of Sequence 2 is found at Position 4 of Sequence 1, A at Position 5 is found at Positions 1, 3 and 6 of Sequence 1, and G at Position 4 found at Position 5 in Sequence 1. For each pair of identities, an offset value is calculated. For example, the R at Position 1 in Sequence 2 matches the R at Position 4, with an offset of 4 - 1 = 3. The R at Position 3 also matches with the R at Position 4 of Sequence 1, with an offset of 1. The G at Position 4 has an offset of 1 with respect to the G in Sequence 1. The A at Position 5 has offset values of -4, -2 and 1. An offset of 1 would have 3 identical residue matches (R, G, and A in both sequences) while other offset values would have one or no matches.

This example describes a lookup table for length *ktup* = 1. Another typical value for the FASTA parameter *ktup* is 2. In this case, FASTA would create a lookup table for pairs of residues, such as R G or G A in the two sequences above.

The algorithm proceeds with a diagonal method, placing each sequence on either the horizontal or vertical axis of a dot-matrix plot. Diagonal lines in the plot indicate identities having the same offset. The score for each diagonal is increased for each identity and decreased for each mismatch. The highest scoring diagonals represent areas of greatest local similarity, and are selected for re-scoring using an amino acid substitution matrix such as PAM-250. The use of a substitution matrix takes into the account the greater likelihood of an amino acid substitution, rather than an insertion or deletion. Matrices like PAM-250 also account for greater likelihood of conserved substitution, where one amino acid replaces another of similar character. The matrix would give a higher score to the replacement of one hydrophobic amino acid with another, for example, and a lower score to replacement of a hydrophobic with a hydrophilic residue. Substitution matrices are defined further in the discussion of the BLAST program.

For each of the highest scoring diagonals, a subregion, or initial region, of maximal similarity is located. If compatible, initial segments may be joined to form a single optimal alignment.

Alignment scores in FASTA are scaled to correct for the length dependence of similarity scoring. Statistical significance of these scores is calculated from the distribution of alignment scores of unrelated sequences [Pearson, 1998]. After parameter estimation for location and scale, similarity

scores for FASTA follow an extreme-value distribution. The statistical significance of length-corrected alignment scores may be calculated from this distribution. The expectation value (e-value) in FASTA provides a measure of statistical significance of alignment scores.

1.2.2 Basic Local Alignment Search Tool (BLAST)

The BLAST family of programs performs sequence similarity searches by finding similar segments between query and subject, and computing the statistical significance of the alignment of these segments. The basic unit of BLAST output is the High Scoring Pair (HSP), or "hit". The following discussion of algorithms and statistics refers to the BLASTP program, which compares protein queries against protein databases [Altschul, et. al., 1997].

The BLAST program begins by creating a list of equal-length high-scoring words between two sequences that meet parameters for score threshold T and word length W. Like FASTA, scores are calculated from substitution matrices, though they are not corrected for length. The BLAST program extends these words and attempts to find segments of maximal cumulative score, or HSPs.

For each HSP, BLAST reports a raw score S calculated from a substitution matrix, and a score normalized by statistical parameters λ and K that describe the "scale" and "location" of the distribution of alignment scores [Pearson, 1998]. These two parameters are calculated from a random model providing

background frequency P_i for each amino acid position in a protein, and a score s_{ij} for alignment of two amino acids using a given substitution matrix. Equation 1.1 defines the expected score for two random amino acids.

$$S = \Sigma P_i P_j s_{ij} \quad \textbf{(1.1)}$$

Normalization of raw scores with parameters λ and K allows score comparison of alignments using different substitution matrices. Normalized scores S′ are reported in bits, and are calculated from the equation:

$$S' = (\lambda S - \ln K)/\ln 2 \quad \textbf{(1.2)}$$

The program reports an expectation value, or e-value, which describes the probability that the bit score S′ occurs by random chance. An e-value of 0.01, for example, says that you can expect to see an equal or higher score by chance in 1 out of 100 alignments. Equation 1.3 defines calculation of the e-value.

$$E = mn/2^{S'} \quad \textbf{(1.3)}$$

The values m and n represent lengths of two protein sequences, and their product the size of the effective search space, or space of all possible solutions. When applied to a protein database, the value n represents the length, or number of residues, in the database. The BLAST program reports those HSPs with e-values less than the default cutoff of 10.

1.2.3 E-value Thresholds

Determination of a threshold or cutoff for e-values is subjective, depending on the investigator's concern for either a false-positive (Type I) or false-negative (Type II) error. Higher e-value thresholds invite more of the former, and lower thresholds more of the latter type of error. Since e-value increases linearly with database size (see Equation 1.3), the choice of smaller databases offers proportionally smaller e-values and greater sensitivity in the search for homologs. Typical e-value cutoffs for small-scale homology searches may be from 0.001 to 0.01, while for large-scale searches involving thousands of sequences, thresholds may typically be set from 10^{-20} to 10^{-6}.

1.2.4 Reference Sequence Project (Refseq)

The Refseq project at the National Center for Biotechnology Information (NCBI) contains a non-redundant set of reference sequences for genomic contigs (overlapping collections of DNA sequences or clones), mRNA and proteins for humans and other organisms including mouse and rat. Curated Refseq records

contain information compiled from multiple sources, so that each record provides current knowledge of known genes. The value of Refseq lies in the curation and review of each record and the non-redundancy of the database. For the implementation of this project, all known drug target protein sequences were retrieved from the Refseq protein database, RefseqP.

1.2.5 Draft Sequence of the Human Genome

Shortly after completion of sequencing and assembly, a draft version of the human genome was made publicly available. The draft is the product of 20 international sequencing centers, and at present is in an 85% finished, highly accurate state. Sequencing work continues, and periodic updates or "freezes" occur frequently. The final version of the human genome sequence is projected for completion in April of 2003. The genome may be searched at public websites including GenBank at the National Center for Biotechnology Information and the University of California Santa Cruz Genome Bioinformatics sites.

The draft genome produced by the International Human Genome Sequencing Consortium employed a strategy known as hierarchical shotgun sequencing. In this method, genomic DNA is fragmented and inserted into a cloning vector, commonly a bacterial artificial chromosome (BAC). The set of cloned DNA fragments represents a genomic library, which are organized into a physical map spanning the entire genome. Selected BAC clones are sequenced with the random shotgun strategy, and clone sequences assembled into the full

genome sequence after filtering for cross-species contaminant data and merging sequence data from overlapping clones [International Human Genome Sequencing Consortium, 2001].

Because the human genome is composed of at least 50% repeated sequence, assembly requires screening for known repetitive sequences. In the genome assembly described here, the RepeatMasker program from Washington University was employed to screen for annotated repetitive sequences and low complexity regions. The sequence output of this program shows a series of N's in place of repetitive sequences and low complexity regions.

Translation of DNA sequence data into protein sequence data may be performed automatically in the following way. Each transcribed three-nucleotide DNA codon is translated into a single amino acid using a codon table. A sliding window produces three different reading frames in both forward and reverse directions by advancing the frame one nucleotide at a time. A six-frame translation results from reading all three frames in each of two directions. To differentiate protein coding from non-coding regions, the procedure finds start codons, which start the process of translation, and stop codons, which terminate translation. A six-frame translation provides DNA sequence translation between two start codons, between a start and a stop codon, and between two stop codons. Functional genes are characterized by the presence of a sequence of nucleotides that are transcribed into at least one start codon and stop codon. Transcription of a gene produces open reading frames (ORFs), frames consisting

of a series of codons without stop codons that may potentially be translated into protein. Figure 1.1 gives a graphical representation of a six-frame translation.

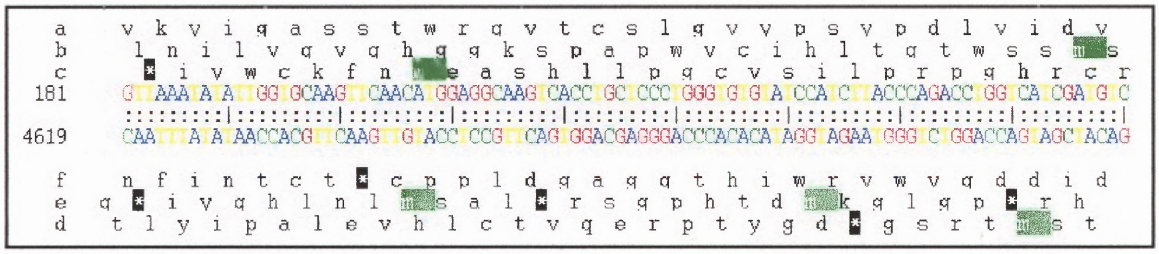

Figure 1.1 Six-frame translation. Stop codons are shown as asterisks and start codons with abbreviation m for methionine. The six reading frames are labeled with a, b and c in the forward direction, and f, e and d in the reverse direction. [Schwager, 2002]

Searching a translation of the genome, rather than genomic DNA itself, is preferable for a number of reasons. Sequence similarity measured by percent sequence identity alone is inferior to similarity determined from substitution matrices such as the PAM-250 matrix, which recognizes conservative substitutions of one amino acid for another with like properties [Wood & Pearson, 1999]. Amino acid sequence searches offer much greater sensitivity than DNA searches as a result of the use of substitution matrices. In addition, searching a translation of the entire genome sequence allows detection of novel genes. A direct search of a protein database will find homologs of already known genes, not as yet unidentified ones.

CHAPTER 2

DEVELOPING *IN-SILICO* PROTEINS

2.1 Problem Statement

With a draft of the human genome now completed, attention has turned to analyzing genes and their protein products, particularly for disease relevance and drug discovery. While newer tools like gene expression arrays and EST databases have identified many gene products as potential targets, they have also inundated pharmaceutical laboratories with a backlog of targets to screen through much more laborious and costly procedures [Federsel, 2001]. A new challenge has arisen in "prioritizing the proteome" [Swindells & Overington, 2002] and in identifying the most likely new drug targets. Computational means can identify and screen potential targets and produce the best candidates for the laboratory, thereby streamlining drug discovery and eliminating much of the labor involved in target validation.

The procedure implemented here identifies genes with target potential, using targets of marketed drugs to start the search. It is based on the assumption that novel targets will have high sequence similarity to established drug receptors.

2.2 Procedure for Creating *In-silico* Proteins - An Overview

Creation of *in-silico* drug target homologs entails these steps:

1. Identify all known targets of marketed drugs.
2. Retrieve accession numbers and sequences of all known targets from a protein database.
3. Query the translated human genome with all target protein sequences.
4. Assemble hits to the translation into *in-silico* proteins.
5. Check *in-silico* protein similarity to the original drug target sequence by querying the protein database in step 2 with *in-silico* sequences.

This procedure for creating *in-silico* proteins is based on one developed by Jeff Yuan at Merck Research Labs. Jeff devised a means for generating *in-silico* gene homologs, or "working genes." Alex Elbrecht recognized the utility of this procedure in generating *in-silico* protein homologs for drug targets or other protein classes, and supervised the work described in the following pages. Bruce Bush, also of Merck Research Labs, provided the genome translation and a Perl script used in steps 3 and 4 above to assemble *in-silico* proteins from FASTA3 hits.

CHAPTER 3

IMPLEMENTATION

3.1 Identifying Known Drug Targets

In the absence of a comprehensive and commercially available list of known drug targets, a list was compiled from *Goodman & Gilman's The Pharmacological Basis of Therapeutics,* the Investigational Drugs Database, and DrugBank (see refs.). Target names extracted from these three resources were used as queries to two protein databases, Swissprot and RefseqP. The compiled list of drug targets is shown in Table A.1.

3.2 Protein Sequence Retrieval

Both Swissprot and RefseqP provide protein annotation, sequence data and links to literature. In many cases, simply evaluating the list of protein names returned for a text query was sufficient to determine which retrieved proteins fit the description for a given drug target. In other cases, a review of the annotation from both databases thinned the number of proteins retrieved. For example, annotation mentioning a protein's role in disease or the fact that it therapeutically bound a drug identified it as the desired receptor.

In many cases, a selected protein had multiple transcript variants and isoforms. When annotation did not designate a specific variant or isoform as a target, all splice variants and isoforms were included.

The final list contained 377 RefseqP accession numbers representing known human therapeutic drug target proteins. Amino acid sequences for each of these 377 accession numbers were retrieved from the RefseqP database.

3.3 Searching the Translated Human Genome

Drug target sequences were queried against a six-frame translation of the April 2002 human genome assembly released by the UCSC Human Genome Project, also known as Golden Path. The assembly was repeat-masked, or screened for repetitive sequences or low complexity regions. Known repetitive sequences and regions of high uncertainty are represented by strings of X's in the translation.

The FASTA3 program was employed to search the translation. When compared to other similarity search programs including BLAST, FASTA3 is, in general, more accurate [Pearson, 1998] and more sensitive to detection of homologs [Jeff Yuan, manuscript in preparation].

3.4 *In-silico* Protein Assembly

Hits to the translated genome, which represent open reading frames, were assembled into *in-silico* proteins if FASTA3 had assigned an e-value less than 1.0. This cut-off e-value was chosen through the experience of Jeff Yuan, who has found relevant FASTA3 hits with e-values slightly less than 1.0. The ORFs were sorted by chromosome, by index number and by orientation (forward or reverse). Regions between ORFs are represented by four dashes (----). The procedure assembled 55,465 *in-silico* proteins from the FASTA3 output. Sample FASTA3 output and *in-silico* protein sequence are shown in Figure B.1.

3.5 Verifying *In-silico* Proteins

Once assembled, all *in-silico* proteins were searched against the RefseqP human protein database using ungapped BLAST set to return only the topmost hit. Since BLAST ranks hits by increasing e-value, the first hit in most cases will have the highest similarity to the query sequence. Comparing the accession numbers of the generating protein and the first hit is a simple and quick task. If the first BLAST hit for an *in-silico* protein was its generating protein, the *in-silico* protein was considered a verified target homolog. Later, a modification of this procedure included reporting of several hits, since identical e-values may appear for transcript variants producing identical proteins. This modification is described further along with results.

3.6 Parsing BLAST reports

Each *in-silico* protein returned a BLAST report, resulting in 55,465 individual reports. In most of these reports, the generating drug target was not the top hit for that *in-silico* protein. Since initially only the top hit is considered valid in this procedure, most of these reports were overlooked in favor of those returning the generating target as the top hit.

A BLAST report parser, written in the Perl programming language, was employed to scroll through each file and compare the RefseqP accession number of an *in-silico* protein's generating target with that of the top hit. If the accession numbers were identical, the parser extracted additional information from the file, including the name of the hit, HSP lengths and e-values, and bit scores. The parser code, which relies on the Bioperl module BPlite, is shown in Section C.1. Sample parsed BLAST output is shown in Section C.2.

3.7 Results

The procedure described above produced 55,465 *in-silico* proteins representing 377 drug targets. Of these 55,465 *in-silico* proteins, 1,203 (2.17%) successfully found their generating drug target as the topmost hit returned from BLAST.

Figure 3.1 compares the number of drug targets with the number verified *in-silico* proteins. Of 377 targets, 54 failed to generate any *in-silico* proteins that in turn found them as first BLAST hit, and 109 targets had exactly one *in-silico*

protein finding self. In most of these one-hit cases, the *in-silico* protein represents the actual target itself. The remaining 214 targets had two or more *in-silico* proteins that were more similar to them than to any other protein in the RefseqP database.

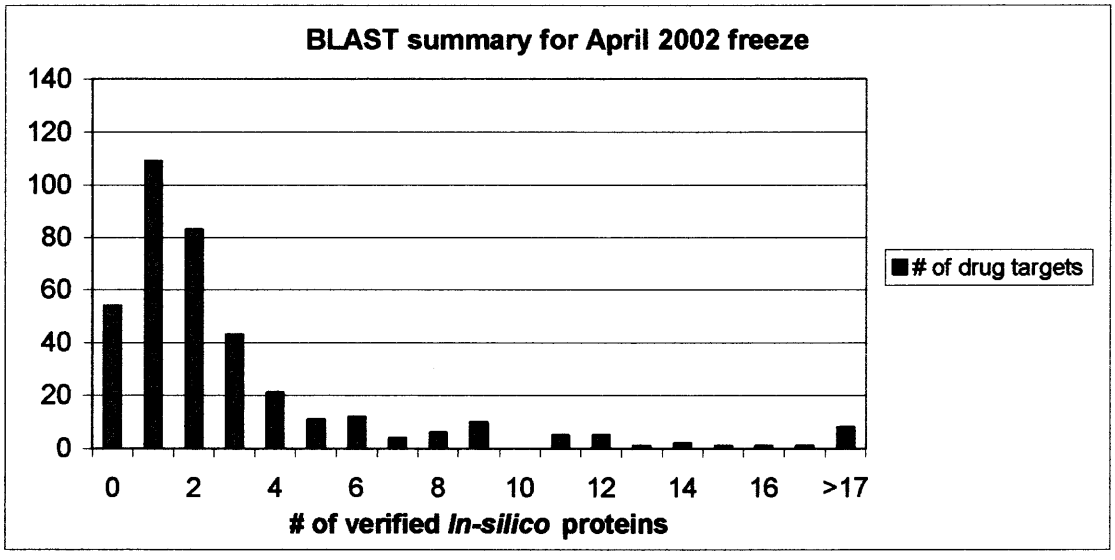

Figure 3.1 BLAST report summary for *in-silico* proteins. Only the first hit was returned for each BLAST report in this summary.

A review of the 54 drug targets with no *in-silico* protein top hits to self revealed that splice variants of many of these 54 appeared first in BLAST reports. In some cases the longest of a set of variants returned the lowest e-value. Since raw alignment scoring is cumulative (Equation 1.1), longer splice variants may return higher scores and lower e-values simply as a function of their length. Figure B.2 shows one example. In other cases, a set of splice variants from a single gene coded for identical proteins. While the target generating an *in-silico*

protein may not have appeared first, it may have appeared further down the list and may have had the same e-value as the first hit listed. An example is shown in Figure B.3.

A review of the 109 drug targets with one *in-silico* protein top hit to self revealed that some of these verified *in-silico* proteins were located on a different chromosome than the drug target that generated them. In most of these cases, as in the case above, BLAST returned the longest of a set of splice variants, or one of several identical proteins as top hit.

In-silico proteins for drug targets having no top hits to self, as well as those having *in-silico* proteins located on different chromosomes than self, were BLASTed again and the top six or twelve hits reported. Listing of the top twelve hits was required only for a set of identical fibroblast growth factor receptors (FGFR2), which have a dozen identical proteins from as many splice variants. Others had a maximum of six splice variants and required reporting of only the top six hits.

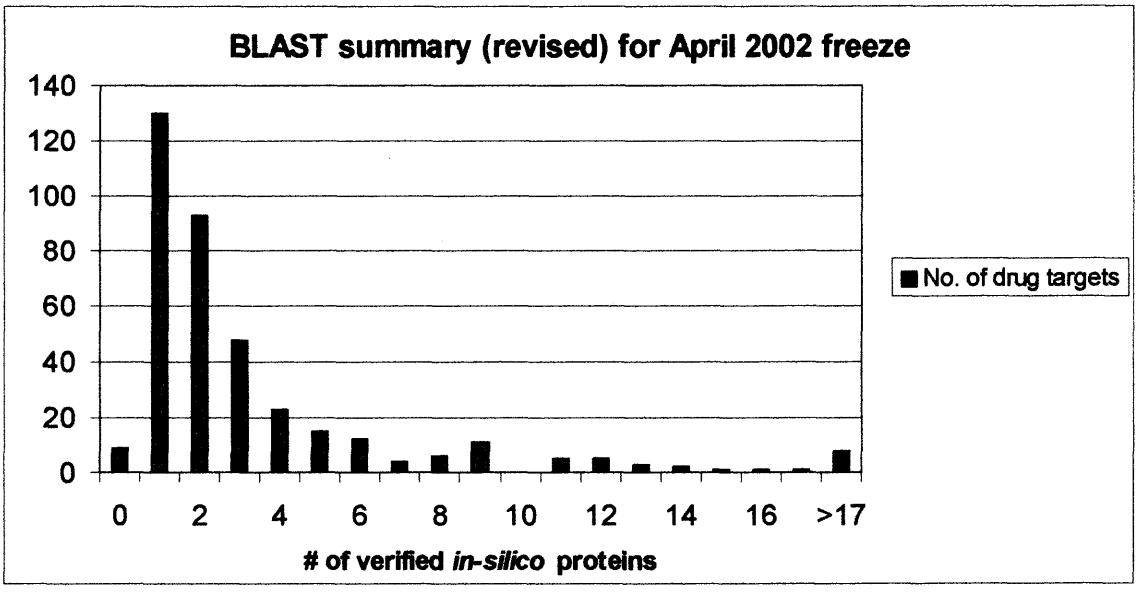

Figure 3.2 Revised BLAST report summary for *in-silico* proteins. These BLAST reports returned either 6 or 12 hits in order to capture equal e-values for multiple splice variants.

After testing the top six or twelve BLAST hits for lowest e-value, nine drug targets having no *in-silico* protein top hits to self remain. Peroxisome proliferative activated receptor (PPAR) gamma, isoform 1 (Refseq accession number NP_619726) demonstrates the BLAST program's assignment of lowest e-value to longest splice variants. In three of the BLAST reports for PPAR gamma, other variants of PPAR including alpha and delta received the lowest e-value. PPAR gamma received higher e-value scores than these variants in spite of the fact that its sequence had fashioned the *in-silico* proteins. The other eight drug targets returned similar results or failed to generate successful *in-silico* proteins for reasons that can only be revealed by a thorough analysis of their sequence and relationships to other proteins.

The number of targets having one *in-silico* protein finding self increased from 109 to 130 as a result of examination of several hits instead of only the first. Most of these *in-silico* proteins represent the sequence of the actual target itself. The remaining 238 drug targets had two or more *in-silico* homologs that identified them as the hit with lowest e-value, some of which may represent novel targets.

CHAPTER 4

CONCLUSIONS

This procedure for generating *in-silico* protein homologs from genomic sequence data is shown to be effective given the following summary of results. Greater than two percent of *in-silico* proteins generated have highest similarity to their generating target as measured by BLAST e-value. A very simple measure of the success of this procedure is the existence of at least one verified *in-silico* protein homolog per drug target. If successful, the procedure should at a minimum produce one *in-silico* homolog representing the sequence of the generating drug target itself. Only 9 of 377 drug targets were unable to generate verified *in-silico* homologs.

Another measure of this procedure's effectiveness is the existence of multiple *in-silico* homologs representing the same drug target. As shown by Figure 3.2, 238 drug targets have more than one verified *in-silico* protein. Some of these *in-silico* proteins may represent novel genomic targets that have as yet not been found or characterized. Thorough analysis and review of these *in-silico* homologs is required in order to locate potential new targets. Once located, other procedures such as microarray expression or EST analysis may be employed for further screening.

Of the 130 drug targets having only one verified *in-silico* homolog, approximately 10% generated *in-silico* homologs located on chromosomes other than their own, indicating an error rate of ~10% in the procedure. Longer splice variants accumulated higher BLAST raw scores and lower e-values than their shorter counterparts, effectively drawing *in-silico* proteins generated by shorter variants nearer to themselves. For these cases, more rigorous algorithms than those used by BLASTP, especially those that correct for query sequence length, may be employed to refine alignments and to find the generating target at its correct location within the genome.

The procedure as detailed requires input of a set of protein accession numbers and some file manipulation. Aside from this, very little user intervention is needed. No decision needs to be made by the user regarding e-value cutoffs, nor any criteria established to determine the validity of a hit. An automatic e-value cutoff of 1.0 for FASTA3 results during *in-silico* protein assembly is liberal enough to include homologous sequences, yet exclusive enough to produce *in-silico* proteins of adequate specificity. No e-value cutoff is required for BLAST results, since only the top-ranking hits are evaluated. The output of the procedure, a set of *in-silico* protein homologs, is a valuable collection that potentially contains new genomic drug targets. Once complete, the set requires analysis and further study in order to determine whether novel targets have been discovered.

The procedure itself may be applied to any group or subgroup of proteins. In upcoming work this procedure will be utilized to create *in-silico* homologs of disease-related proteins, which will also be studied for identification of novel drug targets.

APPENDIX A

KNOWN HUMAN DRUG TARGET PROTEINS

Table A.1 shows the compiled list of 377 human drug target proteins. This list does not include drug metabolizing proteins.

Table A.1 Known Human Drug Target Proteins

Gene Name	RefseqP Accession Number and Description
ABAT	NP_000654 4-aminobutyrate aminotransferase precursor
ABCC5	NP_005679 ATP-binding cassette, sub-family C, member 5
ABL1	NP_005148 v-abl Abelson murine leukemia viral oncogene homolog 1 isoform a
ABL1	NP_009297 v-abl Abelson murine leukemia viral oncogene homolog 1 isoform b
ACAT1	NP_000010 acetyl-Coenzyme A acetyltransferase 1 precursor
ACE	NP_000780 angiotensin I converting enzyme
ACE2	NP_068576 angiotensin I converting enzyme (peptidyl-dipeptidase A) 2, 805 aa.
ACHE	NP_000656 acetylcholinesterase hydrophilic form precursor, 614 aa.
ACHE	NP_056646 acetylcholinesterase PI-linked form precursor Homo sapiens, 617 aa.
ADA	NP_000013 adenosine deaminase
ADORA1	NP_000665 adenosine A1 receptor Homo sapiens, 326 aa.
ADORA2A	NP_000666 adenosine A2a receptor
ADORA2B	NP_000667 adenosine A2b receptor Homo sapiens, 332 aa.
ADORA3	NP_000668 adenosine A3 receptor Homo sapiens, 318 aa.
ADRA1A	NP_000671 alpha-1A-adrenergic receptor, isoform 1
ADRA1A	NP_150645 alpha-1A-adrenergic receptor, isoform 3
ADRA1A	NP_150646 alpha-1A-adrenergic receptor, isoform 2
ADRA1A	NP_150647 alpha-1A-adrenergic receptor, isoform 4
ADRA1B	NP_000670 alpha-1B-adrenergic receptor
ADRA1D	NP_000669 alpha-1D-adrenergic receptor
ADRA2A	NP_000672 alpha-2A-adrenergic receptor
ADRA2B	NP_000673 alpha-2B-adrenergic receptor
ADRA2C	NP_000674 alpha-2C-adrenergic receptor
ADRB1	NP_000675 beta-1-adrenergic receptor [Homo sapiens] - Homo sapiens, 477 aa.
ADRB2	NP_000015 adrenergic, beta-2-, receptor, surface
ADRB3	NP_000016 adrenergic, beta-3-, receptor Homo sapiens, 408 aa.
AGTR1	NP_000676 angiotensin receptor 1
AGTR1	NP_004826 angiotensin receptor 1

AGTR1	NP_033611 angiotensin receptor 1
AGTR1	NP_114038 angiotensin receptor 1
AGTR1	NP_114438 angiotensin receptor 1
AGTR2	NP_000677 angiotensin receptor 2 [Homo sapiens] - Homo sapiens, 363 aa.
AKAP5	NP_004848 A kinase (PRKA) anchor protein 5
ANXA1	NP_000691 annexin I
ANXA5	NP_001145 annexin V
AR	NP_000035 androgen receptor
ATP12A	NP_001667 ATPase, H+/K+ transporting, nongastric, alpha polypeptide
ATP1A1	NP_000692 ATPase, Na+/K+ transporting, alpha 1 polypeptide
ATP1A2	NP_000693 ATPase, Na+/K+ transporting, alpha 2 (+) polypeptide
ATP1A3	NP_000694 ATPase, Na+/K+ transporting, alpha 3 polypeptide,1013 aa.
ATP1B1	NP_001668 ATPase, Na+/K+ transporting, beta 1 polypeptide
ATP1B2	NP_001669 ATPase, Na+/K+ transporting, beta 2 polypeptide
ATP1B3	NP_001670 ATPase, Na+/K+ transporting, beta 3 polypeptide, 279 aa.
ATP4A	NP_000695 ATPase, H+/K+ exchanging, alpha polypeptide
ATP4B	NP_000696 ATPase, H+/K+ exchanging, beta polypeptide
AVPR1A	NP_000697 arginine vasopressin receptor 1A
AVPR1B	NP_000698 arginine vasopressin receptor 1B
AVPR2	NP_000045 arginine vasopressin receptor 2 Homo sapiens, 371 aa.
BZRP	NP_000705 peripheral benzodiazapine receptor
BZRP	NP_009295 peripheral benzodiazapine receptor short form
CA1	NP_001729 carbonic anhydrase I
CA11	NP_001208 carbonic anhydrase XI precursor
CA12	NP_001209 carbonic anhydrase XII precursor
CA14	NP_036245 carbonic anhydrase XIV precursor
CA2	NP_000058 carbonic anhydrase II
CA3	NP_005172 carbonic anhydrase III [Homo sapiens] - Homo sapiens, 260 aa.
CA4	NP_000708 carbonic anhydrase IV precursor
CA5A	NP_001730 carbonic anhydrase VA, mitochondrial precursor
CA5B	NP_009151 carbonic anhydrase VB, mitochondrial precursor
CA6	NP_001206 carbonic anhydrase VI precursor
CA7	NP_005173 carbonic anhydrase VII
CA8	NP_004047 carbonic anhydrase VIII
CA9	NP_001207 carbonic anhydrase IX precursor
CACNA1A	NP_000059 calcium channel, alpha 1A subunit, isoform 1
CACNA1A	NP_075461 calcium channel, alpha 1A subunit, isoform 2
CACNA1C	NP_000710 calcium channel, voltage-dependent, L type, alpha 1C subunit
CCKAR	NP_000721 cholecystokinin A receptor [Homo sapiens] – Homo sapiens, 428 aa.
CD2	NP_001758 CD2 antigen (p50), sheep red blood cell receptor
CD3D	NP_000723 CD3D antigen, delta polypeptide (TiT3 complex), 171 aa.
CD3E	NP_000724 CD3E antigen, epsilon polypeptide (TiT3 complex), 207 aa.
CD3G	NP_000064 CD3G gamma precursor
CD4	NP_000607 CD4 antigen (p55) [Homo sapiens] - Homo sapiens, 458 aa.
CD44	NP_000601 CD44 antigen (homing function and Indian blood group system)
CD8A	NP_001759 CD8 antigen, alpha polypeptide (p32) Homo sapiens, 235 aa.
CD8B1	NP_004922 CD8 antigen, beta polypeptide 1 (p37) Homo sapiens, 210 aa.

CHRM1	NP_000729 cholinergic receptor, muscarinic 1
CHRM2	NP_000730 cholinergic receptor, muscarinic 2
CHRM3	NP_000731 cholinergic receptor, muscarinic 3
CHRM4	NP_000732 cholinergic receptor, muscarinic 4
CHRM5	NP_036257 cholinergic receptor, muscarinic 5
CHRNA5	NP_000736 cholinergic receptor, nicotinic, alpha polypeptide 5
CHRNA7	NP_000737 cholinergic receptor, nicotinic, alpha polypeptide 7 precursor
CHRNB4	NP_000741 cholinergic receptor, nicotinic, beta polypeptide 4
CNP	NP_149124 2',3'-cyclic nucleotide 3' phosphodiesterase
CNR1	NP_001831 central cannabinoid receptor, isoform a
CNR1	NP_057167 central cannabinoid receptor, isoform a
CNR1	NP_149421 central cannabinoid receptor, isoform b
CNR2	NP_001832 cannabinoid receptor 2 (macrophage) Homo sapiens, 360 aa.
COMT	NP_000745 catechol-O-methyltransferase isoform MB-COMT, 271 aa.
COMT	NP_009294 catechol-O-methyltransferase isoform S-COMT, 221 aa.
CSF1R	NP_005202 colony stimulating factor 1 receptor, formerly McDonough feline sarcoma viral (v-fms) oncogene homolog
CSF2RA	NP_006131 colony stimulating factor 2 receptor, alpha, low-affinity (granulocyte-macrophage) 400 aa.
CSF2RB	NP_000386 colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)
CSF3R	NP_000751 colony stimulating factor 3 receptor (granulocyte)
CYP11A	NP_000772 cytochrome P450, subfamily XIA precursor
CYP11B1	NP_000488 cytochrome P450, subfamily XIB (steroid 11-beta-hydroxylase), polypeptide 1 precursor
CYP11B2	NP_000489 cytochrome P450, subfamily XIB polypeptide 2 precursor
CYP17	NP_000093 cytochrome P450, subfamily XVII polypeptide
CYP19	NP_000094 cytochrome P450, subfamily XIX polyppetide
CYP19	NP_112503 cytochrome P450, subfamily XIX polypeptide
CYSLT1	NP_006630 cysteinyl leukotriene receptor 1 Homo sapiens, 337 aa.
DDC	NP_000781 dopa decarboxylase (aromatic L-amino acid decarboxylase)
DHFR	NP_000782 dihydrofolate reductase Homo sapiens, 187 aa.
DRD1	NP_000785 dopamine receptor D1 Homo sapiens, 446 aa.
DRD2	NP_000786 dopamine receptor D2, isoform long Homo sapiens, 443 aa.
DRD2	NP_057658 dopamine receptor D2, isoform short Homo sapiens, 414 aa.
DRD3	NP_000787 dopamine receptor D3, isoform a Homo sapiens, 400 aa.
DRD3	NP_387507 dopamine receptor D3, isoform b Homo sapiens, 208 aa.
DRD3	NP_387508 dopamine receptor D3, isoform c Homo sapiens, 327 aa.
DRD3	NP_387509 dopamine receptor D3, isoform d Homo sapiens, 299 aa.
DRD3	NP_387512 dopamine receptor D3, isoform e Homo sapiens, 342 aa.
DRD4	NP_000788 dopamine receptor D4
EDNRA	NP_001948 endothelin receptor type A Homo sapiens, 427 aa.
EPOR	NP_000112 erythropoietin receptor precursor Homo sapiens, 508 aa.
ERBB2	NP_004439 v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
ESR1	NP_000116 estrogen receptor 1
ESR2	NP_001428 estrogen receptor 2 (ER beta) Homo sapiens, 530 aa.
F10	NP_000495 coagulation factor X precursor

F2	NP_000497 coagulation factor II precursor
F9	NP_000124 coagulation factor IX
FGFR2	NP_000132 fibroblast growth factor receptor 2, isoform 1 precursor
FGFR2	NP_075418 fibroblast growth factor receptor 2, isoform 11 precursor
FGFR2	NP_075258 fibroblast growth factor receptor 2, isoform 2 precursor
FGFR2	NP_075259 fibroblast growth factor receptor 2, isoform 3 precursor
FGFR2	NP_075260 fibroblast growth factor receptor 2, isoform 4 precursor
FGFR2	NP_075261 fibroblast growth factor receptor 2, isoform 5 precursor
FGFR2	NP_075262 fibroblast growth factor receptor 2, isoform 6 precursor
FGFR2	NP_075263 fibroblast growth factor receptor 2, isoform 7 precursor
FGFR2	NP_075264 fibroblast growth factor receptor 2, isoform 8 precursor
FGFR2	NP_075265 fibroblast growth factor receptor 2, isoform 9 precursor
FGFR2	NP_075417 fibroblast growth factor receptor 2, isoform 10 precursor
FGFR2	NP_075419 fibroblast growth factor receptor 2, isoform 12 precursor
FGFR2	NP_075420 fibroblast growth factor receptor 2, isoform 13 precursor
FLT1	NP_002010 fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor) Homo sapiens, 1338 aa.
FLT4	NP_002011 fms-related tyrosine kinase 4
FOLR1	NP_000793 folate receptor 1 (adult) Homo sapiens, 257 aa.
FOLR1	NP_057943 folate receptor 1 precursor Homo sapiens, 257 aa.
FOLR1	NP_057936 folate receptor 1 precursor Homo sapiens, 257 aa.
FOLR1	NP_057937 folate receptor 1 precursor Homo sapiens, 257 aa.
FOLR1	NP_057941 folate receptor 1 precursor Homo sapiens, 257 aa.
FOLR1	NP_057942 folate receptor 1 precursor Homo sapiens, 257 aa.
FOLR2	NP_000794 folate receptor 2 precursor Homo sapiens, 255 aa.
FOLR3	NP_000795 folate receptor 3 precursor Homo sapiens, 243 aa.
FRAP1	NP_004949 FK506 binding protein 12-rapamycin associated protein 1
FXYD2	NP_001671 FXYD domain-containing ion transport regulator 2, isoform 1
FXYD2	NP_067614 FXYD domain-containing ion transport regulator 2, isoform 2
G2AN	NP_055425 alpha glucosidase II alpha subunit
GAA	NP_000143 acid alpha-glucosidase preproprotein
GABBR1	NP_001461 gamma-aminobutyric acid (GABA) B receptor 1 isoform a precursor
GABBR1	NP_068703 gamma-aminobutyric acid (GABA) B receptor 1 isoform b precursor
GABBR1	NP_068704 gamma-aminobutyric acid (GABA) B receptor 1 isoform c precursor
GABBR1	NP_068705 gamma-aminobutyric acid (GABA) B receptor 1 isoform d precursor
GABRA1	NP_000797 gamma-aminobutyric acid (GABA) A receptor, alpha 1 precursor
GABRA2	NP_000798 gamma-aminobutyric acid A receptor, alpha 2 precursor, 451 aa.
GABRA3	NP_000799 gamma-aminobutyric acid A receptor, alpha 3 precursor
GABRA4	NP_000800 gamma-aminobutyric acid A receptor, alpha 4 precursor, 554 aa.
GABRA5	NP_000801 gamma-aminobutyric acid (GABA) A receptor, alpha 5 precursor
GABRA6	NP_000802 gamma-aminobutyric acid A receptor, alpha 6 precursor, 453 aa.
GABRB1	NP_000803 gamma-aminobutyric acid (GABA) A receptor, beta 1 precursor
GABRB2	NP_000804 gamma-aminobutyric acid (GABA) A receptor, beta 2, isoform 2
GABRB3	NP_000805 gamma-aminobutyric acid (GABA) A receptor, beta 3, isoform 1 precursor Homo sapiens, 473 aa.
GABRD	NP_000806 gamma-aminobutyric acid (GABA) A receptor, delta, 452 aa.
GABRE	NP_004952 gamma-aminobutyric acid (GABA) A receptor, epsilon, isoform 1 precursor Homo sapiens, 506 aa.

GABRE	NP_068819 gamma-aminobutyric acid (GABA) A receptor, epsilon, isoform 2
GABRE	NP_068822 gamma-aminobutyric acid (GABA) A receptor, epsilon, isoform 3
GABRE	NP_068830 gamma-aminobutyric acid (GABA) A receptor, epsilon, isoform 2
GABRG2	NP_000807 gamma-aminobutyric acid A receptor, gamma 2 precursor, 467 aa.
GABRG3	NP_150092 gamma-aminobutyric acid (GABA) A receptor, gamma 3, 467 aa.
GABRP	NP_055026 gamma-aminobutyric acid (GABA) A receptor, pi, 440 aa.
GNRHR	NP_000397 gonadotropin-releasing hormone receptor
GNRHR2	NP_476504 gonadotropin-releasing hormone (type 2) receptor 2, 292 aa.
GPR38	NP_001498 G protein-coupled receptor 38 Homo sapiens, 412 aa.
GRIA1	NP_000818 glutamate receptor, ionotropic, AMPA 1 precursor
GRIA2	NP_000817 glutamate receptor, ionotropic, AMPA 2 precursor
GRIA3	NP_000819 glutamate receptor, ionotrophic, AMPA 3 isoform flip
GRIA3	NP_015564 glutamate receptor, ionotrophic, AMPA 3 isoform flip
GRIA4	NP_000820 glutamate receptor, ionotrophic
GRIK1	NP_000821 glutamate receptor, ionotropic, kainate 1
GRIK2	NP_068775 glutamate receptor, ionotropic, kainate 2 Homo sapiens, 908 aa.
GRIK3	NP_000822 glutamate receptor, ionotropic, kainate 3 Homo sapiens, 919 aa.
GRIK4	NP_055434 glutamate receptor, ionotropic, kainate 4
GRIK5	NP_002079 glutamate receptor, ionotropic, kainate 5 Homo sapiens, 981 aa.
GRIN1	NP_000823 NMDA receptor 1, isoform NR1-1 precursor
GRIN1	NP_015566 NMDA receptor 1, isoform NR1-3 precursor
GRIN1	NP_067544 NMDA receptor 1, isoform NR1-2 precursor
GRIN2A	NP_000824 N-methyl-D-aspartate receptor subunit 2A precursor, 1464 aa.
GRIN2B	NP_000825 N-methyl-D-aspartate receptor subunit 2B precursor, 1484 aa.
GRIN2C	NP_000826 N-methyl-D-aspartate receptor subunit 2C precursor, 1236 aa.
GRIN2D	NP_000827 N-methyl-D-aspartate receptor subunit 2D precursor
GRIN3A	NP_597702 glutamate receptor, ionotropic, N-methyl-D-aspartate 3A, 1115 aa.
GUCY1A2	NP_000846 guanylate cyclase 1, soluble, alpha 2 Homo sapiens, 732 aa.
GUCY1A3	NP_000847 guanylate cyclase 1, soluble, alpha 3 Homo sapiens, 717 aa.
GUCY1B2	NP_004120 guanylate cyclase 1, soluble, beta 2 Homo sapiens, 617 aa.
GUCY1B3	NP_000848 guanylate cyclase 1, soluble, beta 3 Homo sapiens, 619 aa.
GUCY2C	NP_004954 guanylate cyclase 2C (heat stable enterotoxin receptor)
HKE2	NP_055075 HLA class II region expressed gene KE2
HKE4	NP_008910 HLA class II region expressed gene KE4
HLA-A	NP_002107 major histocompatibility complex, class I, A precursor
HLA-B	NP_005505 major histocompatibility complex, class I, B precursor
HLA-C	NP_002108 major histocompatibility complex, class I, C precursor
HLA-DMA	NP_006111 major histocompatibility complex, class II, DM alpha precursor
HLA-DOA	NP_002110 major histocompatibility complex, class II, DO alpha
HLA-DOB	NP_002111 major histocompatibility complex, class II, DO beta precursor
HLA-DPA1	NP_291032 major histocompatibility complex, class II, DP alpha 1
HLA-DPB1	NP_002112 major histocompatibility complex, class II, DP beta 1 precursor
HLA-DRA	NP_061984 major histocompatibility complex, class II, DR alpha precursor
HLA-DRB1	NP_002115 major histocompatibility complex, class II, DR beta 1 precursor
HLA-DRB5	NP_002116 major histocompatibility complex, class II, DR beta 5 precursor
HLA-E	NP_005507 major histocompatibility complex, class I, E precursor
HLA-G	NP_002118 major histocompatibility complex, class I, G precursor

HMGCR	NP_000850 3-hydroxy-3-methylglutaryl-Coenzyme A reductase, 888 aa.
HRH1	NP_000852 histamine receptor H1
HRH2	NP_071640 histamine receptor H2
HRH3	NP_009163 histamine receptor H3
HTR1A	NP_000515 5-hydroxytryptamine (serotonin) receptor 1A Homo sapiens, 421 aa.
HTR1D	NP_000855 5-hydroxytryptamine (serotonin) receptor 1D Homo sapiens, 377 aa.
HTR2A	NP_000612 5-hydroxytryptamine (serotonin) receptor 2A Homo sapiens, 471 aa.
HTR2B	NP_000858 5-hydroxytryptamine (serotonin) receptor 2B Homo sapiens, 481 aa.
HTR2C	NP_000859 5-hydroxytryptamine (serotonin) receptor 2C Homo sapiens, 458 aa.
HTR3A	NP_000860 5-hydroxytryptamine (serotonin) receptor 3A
HTR3B	NP_006019 5-hydroxytryptamine (serotonin) receptor 3B
HTR3C	NP_570126 5-hydroxytryptamine receptor 3 subunit C Homo sapiens, 447 aa.
HTR4	NP_000861 5-hydroxytryptamine (serotonin) receptor 4 Homo sapiens, 388 aa.
IFNGR1	NP_000407 interferon gamma receptor 1
IL2RA	NP_000408 interleukin 2 receptor, alpha chain precursor Homo sapiens, 272 aa.
IL4R	NP_000409 interleukin 4 receptor precursor
IL6R	NP_000556 interleukin 6 receptor Homo sapiens, 468 aa.
IMPDH1	NP_000874 IMP (inosine monophosphate) dehydrogenase 1
IMPDH2	NP_000875 IMP (inosine monophosphate) dehydrogenase 2
INSR	NP_000199 insulin receptor Homo sapiens, 1382 aa.
ITGA2B	NP_000410 integrin alpha 2b precursor Homo sapiens, 1039 aa.
ITGAL	NP_002200 integrin alpha L precursor
ITGB2	NP_000202 integrin beta chain, beta 2 precursor
ITGB3	NP_000203 integrin beta chain, beta 3 precursor
K-ALPHA-1	NP_006073 tubulin, alpha, ubiquitous Homo sapiens, 451 aa.
KCNJ1	NP_000211 potassium inwardly-rectifying channel, subfamily J, member 1
KCNJ5	NP_000881 potassium inwardly-rectifying channel, subfamily J, member 5
KCNJ6	NP_002231 potassium inwardly-rectifying channel, subfamily J, member 6
KCNJ8	NP_004973 potassium inwardly-rectifying channel, subfamily J, member 8
LIPE	NP_005348 hormone-sensitive lipase
MAOA	NP_000231 monoamine oxidase A Homo sapiens, 527 aa.
MAOB	NP_000889 monoamine oxidase B Homo sapiens, 520 aa.
MAPK14	NP_001306 mitogen-activated protein kinase 14, isoform 1
MAPK14	NP_620581 mitogen-activated protein kinase 14, isoform 2
MAPK14	NP_620582 mitogen-activated protein kinase 14, isoform 3
MAPK14	NP_620583 mitogen-activated protein kinase 14, isoform 4
MARS	NP_004981 methionine-tRNA synthetase
MC2R	NP_000520 melanocortin 2 receptor
MME	NP_000893 membrane metallo-endopeptidase
MME	NP_009218 membrane metallo-endopeptidase
MME	NP_009219 membrane metallo-endopeptidase
MME	NP_009220 membrane metallo-endopeptidase
MTP	NP_000244 microsomal triglyceride transfer protein large subunit precursor
NQO1	NP_000894 NAD(P)H menadione oxidoreductase 1, dioxin-inducible
NR1D1	NP_068370 nuclear receptor subfamily 1, group D, member 1
NR3C1	NP_000167 nuclear receptor subfamily 3, group C, member 1
NR3C2	NP_000892 nuclear receptor subfamily 3, group C, member 2

OPRD1	NP_000902 opioid receptor, delta 1 Homo sapiens, 372 aa.
OPRK1	NP_000903 opioid receptor, kappa 1
OPRM1	NP_000905 opioid receptor, mu 1 Homo sapiens, 400 aa.
OXTR	NP_000907 oxytocin receptor Homo sapiens, 389 aa.
PDE11A	NP_058649 phosphodiesterase 11A
PDE2A	NP_002590 phosphodiesterase 2A, cGMP-stimulated
PDE3A	NP_000912 phosphodiesterase 3A, cGMP-inhibited Homo sapiens, 1141 aa.
PDE4A	NP_006193 phosphodiesterase 4A, cAMP-specific (phosphodiesterase E2 dunce homolog, Drosophila)
PDE4B	NP_002591 phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila)
PDE4C	NP_000914 phosphodiesterase 4C, cAMP-specific (phosphodiesterase E1 dunce homolog, Drosophila)
PDE4D	NP_006194 phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)
PDE5A	NP_001074 phosphodiesterase 5A, isoform 1
PDE5A	NP_236914 phosphodiesterase 5A, isoform 2
PDE5A	NP_237223 phosphodiesterase 5A, isoform 4
PDE5A	NP_246273 phosphodiesterase 5A, isoform 3
PDE9A	NP_002597 phosphodiesterase 9A
PDGFRB	NP_002600 platelet-derived growth factor receptor beta precursor
PFKFB1	NP_002616 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1
PFKFB2	NP_006203 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2
PFKFB3	NP_004557 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3, 520 aa.
PFKFB4	NP_004558 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4, 469 aa.
PGR	NP_000917 progesterone receptor Homo sapiens, 933 aa.
PLA2G10	NP_003552 phospholipase A2, group X Homo sapiens, 165 aa.
PLA2G12	NP_110448 group XII secreted phospholipase A2 Homo sapiens, 189 aa.
PLA2G13	NP_115951 group XIII secreted phospholipase A2 Homo sapiens, 194 aa.
PLA2G1B	NP_000919 phospholipase A2, group IB (pancreas) Homo sapiens, 148 aa.
PLA2G2A	NP_000291 phospholipase A2, group IIA (platelets, synovial fluid), 144 aa.
PLA2G2D	NP_036532 phospholipase A2, group IID
PLA2G2E	NP_055404 phospholipase A2, group IIE Homo sapiens, 142 aa.
PLA2G2F	NP_073730 phospholipase A2, group IIF Homo sapiens, 168 aa.
PLA2G3	NP_056530 group III secreted phospholipase A2 Homo sapiens, 509 aa.
PLA2G4B	NP_005081 phospholipase A2, group IVB (cytosolic) Homo sapiens, 1012 aa.
PLA2G4C	NP_003697 phospholipase A2, group IVC (cytosolic, calcium-independent)
PLA2G5	NP_000920 phospholipase A2, group V Homo sapiens, 138 aa.
PLA2G6	NP_003551 phospholipase A2, group VI (cytosolic, calcium-independent), 806 aa.
PLA2G7	NP_005075 phospholipase A2, group VII (platelet-activating factor acetylhydrolase, plasma)
PLAUR	NP_002650 plasminogen activator, urokinase receptor Homo sapiens, 335 aa.
PLG	NP_000292 plasminogen Homo sapiens, 810 aa.
POLD4	NP_066996 polymerase (DNA-directed), delta 4
POLE3	NP_059139 polymerase (DNA directed), epsilon 3 (p17 subunit)
POLE4	NP_063949 polymerase (DNA-directed), epsilon 4 (p12 subunit)
POLL	NP_037406 polymerase (DNA directed), lambda
PPARA	NP_005027 peroxisome proliferative activated receptor, alpha, 468 aa.

PPARG	NP_005028 peroxisome proliferative activated receptor gamma, isoform 1
PPARG	NP_056953 peroxisome proliferative activated receptor gamma, isoform 2
PPARG	NP_619725 peroxisome proliferative activated receptor gamma, isoform 1
PPARG	NP_619726 peroxisome proliferative activated receptor gamma, isoform 1
PPP3CA	NP_000935 protein phosphatase 3 (formerly 2B), catalytic subunit, alpha isoform (calcineurin A alpha)
PPP3CB	NP_066955 protein phosphatase 3 (formerly 2B), catalytic subunit, beta isoform (calcineurin A beta)
PPP3CC	NP_005596 protein phosphatase 3 (formerly 2B), catalytic subunit, gamma isoform (calcineurin A gamma)
PPP3R1	NP_000936 protein phosphatase 3 (formerly 2B), regulatory subunit B (19kD), alpha isoform (calcineurin B, type I)
PRKCZ	NP_002735 protein kinase C, zeta Homo sapiens, 592 aa.
PTGER1	NP_000946 prostaglandin E receptor 1 (subtype EP1), 42kD
PTGER2	NP_000947 prostaglandin E receptor 2 (subtype EP2), 53kD
PTGER3	NP_000948 prostaglandin E receptor 3 (subtype EP3)
PTGER4	NP_000949 prostaglandin E receptor 4 (subtype EP4) Homo sapiens, 488 aa.
PTGFR	NP_000950 prostaglandin F receptor (FP), 359 aa.
PTGS1	NP_000953 prostaglandin-endoperoxide synthase 1, isoform 1 precursor
PTGS1	NP_542158 prostaglandin-endoperoxide synthase 1, isoform 2 precursor
PTGS2	NP_000954 prostaglandin-endoperoxide synthase 2 precursor
PTPRC	NP_002829 protein tyrosine phosphatase, receptor type, C, isoform 1 precursor
PTPRC	NP_563578 protein tyrosine phosphatase, receptor type, C, isoform 2 precursor
PTPRC	NP_563579 protein tyrosine phosphatase, receptor type, C, isoform 3 precursor
PTPRC	NP_563580 protein tyrosine phosphatase, receptor type, C, isoform 4
RARA	NP_000955 retinoic acid receptor, alpha
RARB	NP_000956 retinoic acid receptor, beta isoform 1
RARG	NP_000957 retinoic acid receptor, gamma
RRM1	NP_001024 ribonucleoside-diphosphate reductase M1 chain
RRM2	NP_001025 ribonucleotide reductase M2 polypeptide Homo sapiens, 389 aa.
RXRA	NP_002948 retinoid X receptor, alpha Homo sapiens, 462 aa.
RXRB	NP_068811 retinoid X receptor, beta Homo sapiens, 533 aa.
RXRG	NP_008848 retinoid X receptor, gamma Homo sapiens, 463 aa.
SLC12A1	NP_000329 sodium potassium chloride cotransporter 2
SLC12A3	NP_000330 solute carrier family 12 (sodium/chloride transporters), member 3
SLC6A1	NP_003033 solute carrier family 6 (neurotransmitter transporter, GABA), member 1
SLC6A12	NP_003035 solute carrier family 6 (neurotransmitter transporter, betaine/GABA), member 12
SLC6A2	NP_001034 solute carrier family 6 (neurotransmitter transporter, noradrenalin), member 2
SLC6A3	NP_001035 solute carrier family 6 (neurotransmitter transporter, dopamine), member 3
SLC6A4	NP_001036 solute carrier family 6 (neurotransmitter transporter, serotonin), member 4
SLC8A3	NP_150287 solute carrier family 8 (sodium-calcium exchanger), member 3, isoform A precursor
SLC8A3	NP_489479 solute carrier family 8 (sodium-calcium exchanger), member 3, isoform B precursor
SLC9A3R2	NP_004776 solute carrier family 9 isoform 3 regulator 2

SOAT1	NP_003092 sterol O-acyltransferase (acyl-Coenzyme A: cholesterol acyltransferase) 1 Homo sapiens, 550 aa.
SRD5A1	NP_001038 steroid-5-alpha-reductase, alpha polypeptide 1 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 1)
SRD5A2	NP_000339 3-oxo-5 alpha-steroid 4-dehydrogenase 2
SSTR2	NP_001041 somatostatin receptor 2 Homo sapiens, 369 aa.
SSTR3	NP_001042 somatostatin receptor 3 Homo sapiens, 418 aa.
SSTR4	NP_001043 somatostatin receptor 4 Homo sapiens, 388 aa.
SSTR5	NP_001044 somatostatin receptor 5 Homo sapiens, 364 aa.
TBXA2R	NP_001051 thromboxane A2 receptor Homo sapiens, 343 aa.
TBXAS1	NP_001052 thromboxane A synthase 1 (platelet, cytochrome P450, subfamily V), isoform TXS-I
TBXAS1	NP_112246 thromboxane A synthase 1 (platelet, cytochrome P450, subfamily V), isoform TXS-II
TH	NP_000351 tyrosine hydroxylase Homo sapiens, 497 aa.
THRA	NP_003241 thyroid hormone receptor, alpha (avian erythroblastic leukemia viral (v-erb-a) oncogene homolog)
THRB	NP_000452 thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian)
TNF	NP_000585 tumor necrosis factor (cachectin)
TOP1	NP_003277 DNA topoisomerase I
TOP2A	NP_001058 DNA topoisomerase II, alpha isozyme
TOP2B	NP_001059 DNA topoisomerase II, beta isozyme
TPO	NP_000538 thyroid peroxidase Homo sapiens, 933 aa.
TRPV1	NP_061197 transient receptor potential cation channel, subfamily V, member 1
TRPV1	NP_542435 transient receptor potential cation channel, subfamily V, member 1
TRPV1	NP_542436 transient receptor potential cation channel, subfamily V, member 1
TRPV1	NP_542437 transient receptor potential cation channel, subfamily V, member 1
TUBA1	NP_005991 tubulin, alpha 1
TUBA2	NP_524575 tubulin, alpha 2 isoform 2 Homo sapiens, 418 aa.
TUBA2	NP_005992 tubulin, alpha 2 isoform 1 Homo sapiens, 450 aa.
TUBA3	NP_006000 tubulin, alpha 3
TUBA6	NP_116093 tubulin alpha 6 Homo sapiens, 449 aa.
TUBAL2	NP_061816 tubulin, alpha-like 2
TUBB	NP_001060 tubulin, beta polypeptide Homo sapiens, 445 aa.
TUBB1	NP_110400 beta tubulin 1, class VI Homo sapiens, 451 aa.
TUBB2	NP_006079 tubulin, beta, 2 Homo sapiens, 445 aa.
TUBB4	NP_006077 tubulin, beta, 4 Homo sapiens, 450 aa.
TUBB4Q	NP_064424 tubulin, beta polypeptide 4, member Q Homo sapiens, 434 aa.
TUBB5	NP_006078 tubulin, beta, 5 Homo sapiens, 444 aa.
TUBG1	NP_001061 tubulin, gamma 1
TUBG2	NP_057521 tubulin, gamma 2 Homo sapiens, 451 aa.
TYMS	NP_001062 thymidylate synthetase
XDH	NP_000370 xanthene dehydrogenase

APPENDIX B

FASTA3 AND BLASTP OUTPUT

Appendix B gives examples of BLASTP and FASTA3 program outputs and *in-silico* protein sequences.

```
FASTA (3.39 May 2001) function [optimized, BL50 matrix (15:-5)] ktup: 2
The best scores are:                          opt  bits  E(113355542)
1NT_004610_3@000353562-000353861    ( 100) 352  121   7.8e-25
1NT_004610_1@000358149-000358322    (  58) 315  109   1.8e-21
1NT_004610_3@000357405-000357569    (  55) 293  101   2.5e-19
1NT_004610_4@000384153-000383965    (  63) 238   83   7.5e-14
1NT_004610_1@000410916-000411224    ( 103) 204   72   2.9e-10
1NT_004610_5@000246361-000246011    ( 117) 192   68   5e-09
1NT_004610_5@000189913-000189770    (  48) 170   61   2.7e-07
      :     :     :     :           :     :     :     :
      :     :     :     :           :     :     :     :
>1NT_004610F            8.84e-76    7.8e-25
000230653-000230919 000353562-000353861 000357405-000357569
000358149-000358322 000410916-000411224 000412154-000412399
KLTSVFPSPPPPPRYLFPLVSDRGDFCQEAACDCDREAAHCSFDNLGTYNKTMCNYPSFL
CENWFLSAECHPFSLALSGTEAFFPPPAP----GRKGRRCWAVGLKQGWAAWVSPGDMGY
LPGVPAVQGGLLDLKSMIEKVTGKNALTNYGFYGCYCGWGGRGTPKDGTDW----VSCTD
RCCWAHDHCYGRLEEKGCNIRTQSYKYRFAWGVVTCGKAGASRSGHWKST----DWSMLF
PPSFCSWCPFAEPGPFCHVNLCACDRKLVYCLKRNLRSYNPQYQYFPNILCS----DSQD
ERGWERPAPGQGLRAPPGSLAVLSTAHGSLLNLKAMVEAVTGRSAILSFVGYGCYCGLGG
RGQPKDEVDW----AGGRWAGPAMPVHSPLPLPRCCHAHDCCYQELFDQGCHPYVDHYDH
TIENNTEIVCSESLPCHLGPQRREWLALPACAGMVD
```

Figure B.1 FASTA3 output and assembled *in-silico* protein sequence for phospholipase A2, group V (Refseq acc. no. NP_000920).

```
NP_246273 phosphodiesterase 5A, isoform 3
Length = 823
In-silico protein query:
Query= 4NT_006302R       0    6e-120    003622884-003622216
003612134-003612024 003582738-003582511 003580932-003580816
003578492-003578211 003575885-003575817 003569384-003569070
(1299 letters)
Database: /lhome/bushbr/databases/refdb/Hs.aa
          15,001 sequences; 7,710,218 total letters
Sequences producing significant alignments:
                                          Score      E
                                          (bits)     Value      N
NP_237223 phosphodiesterase 5A,           488        0.0        19
          isoform 4; cGMP-binding…

>NP_237223 phosphodiesterase 5A, isoform 4; cGMP-binding cGMP-specific
3',5'-cyclic nucleotide phosphodiesterase [Hs] - Hs, 865 aa.
Length = 865
```

Figure B.2 Sample BLASTP output for an *in-silico* protein sequence derived from phosphodiesterase 5A, isoform 3 (Refseq acc. no. NP_246273). This protein did not generate any verified *in-silico* homologs. Its longer isoform, isoform 4, received lowest e-value and was returned as the top BLAST hit.

Sequences producing significant alignments:	Score (bits)	e-value	N
NP_002111 major histocompatibility complex,class II,DO bet...	41	0.001	1
NP_064455 immunoglobulin lambda-like polypeptide 1;imm...	40	0.001	2
NP_072049 major histocompatibility complex,class II,DR bet...	41	0.001	1
NP_002115 major histocompatibility complex,class II,DR bet...	41	0.001	1
NP_002113 major histocompatibility complex,class II,DQ alp...	40	0.002	1
NP_002112 major histocompatibility complex,class II,DP bet...	39	0.005	1

Figure B.3 Sample BLASTP output for an *in-silico* protein sequence derived from major histocompatibility complex, class II, DR beta 1 precursor (Refseq acc. no. NP_002115). The generating target appears fourth in the list of BLAST hits, and has identical e-value and bit score to the first hit in the list.

APPENDIX C

BLAST PARSER CODE AND SAMPLE OUTPUT

Appendix C shows Perl code for the BLAST parser employed to retrieve data from BLASTP output files. Sample BLAST parser output follows.

Section C.1 Perl code for BLAST parser.

```
#!/usr/bin/perl -w

# This script uses the Bioperl BPlite module to parse BLAST output. It
# takes a file of accession numbers, opens a directory associated with
# each, reads each BLAST file in the directory and creates a single
# file of parsed output for all files and directories.

# Sue McClatchy
# Sept. 15, 2002

use strict;
use Bio::Tools::BPlite;

my $blastdir = "";
my $totalselfhits = 0;
my $totalwpcount = 0;
my %tgtsuccess = ();
my %wpsuccess = ();
my %wpidhash = ();

# Open the accession number file and read.
open (IN, "rspaccs3.txt") || die("Cannot open file\n");
while (my $line = <IN>) {
    chomp($line);
    if ($line =~ /^(NP_\d+)/) {
        $blastdir = $1;
```

```
# Open the directory associated with an accession number and grab all
# files.
    my @hitfiles = glob("$blastdir/*.Hsrsp.blast");
    my $id = "";
    my $wpid = "";
    my $nohit = 0;
    my $selfhitcount = 0;
    my $wpcount = 0;

    LOOP: foreach my $rsphitfile (@hitfiles){
        my @hspqstart = ();
        my @hspqend = ();
        my @hsphstart = ();
        my @hsphend = ();
        $wpcount++;

# For each file, create a new object.
        # Read in input blast file and create new report:
        my $report = new Bio::Tools::BPlite(-file => $rsphitfile);

        # Get working protein ID number:
        open (BLAST, "$rsphitfile") || die("Cannot open file -
        $rsphitfile\n");
        while (my $fline = <BLAST>) {
            chomp($fline);
            unless ($fline) { next; }
            if ($fline =~ /^Query=\s(\w+)(\s+|\S+)/) {
                $wpid = $1;
                $wpidhash{$wpid} = 1;
                }
                if($fline =~ /\*\*\*\*\*\sNo\shits\sfound\s\*\*\*\*\*/)
                    {
                    $nohit++;
                    next LOOP;
                    }
            else { next; }
            }
```

```
# Read in ONLY the first hit in the BLAST report:
my $sbjct = $report->nextSbjct;

# Parse the header line ($sbjct->name) and extract the
refseqp accession number:
if ($sbjct->name =~ /^(NP_\d+)(\s+|\S+)/) {
      $id = $1;

# Compare the id number of the first hit with that of the
# directory. If they are the same, print out the hit name,
# accession number, and the accession number of the
# directory.

    if ($id eq $blastdir) {
        $selfhitcount++;
        $tgtsuccess{$id} = 1;
        $wpsuccess{$wpid} = 1;
      # access the hit name
        print "Hit: " ,$sbjct->name, "\n";
        print "Hit accession no:\t\t" ,$id, "\n";
        print "Working protein created from:\t"
      ,$blastdir,"\n";

      # Place the query and hit starting and ending numbers
      # for each hsp in two arrays. Sort them in ascending
      # order.
        while(my $hsp = $sbjct->nextHSP) {
            # gets the next HSP
            print "HSP: ",$hsp->query->start,"\t"
            ,$hsp->query->end,"\t%ID: "
            ,$hsp->percent,"\tscore: ",$hsp->score,
            "\tbits: ",$hsp->bits,"\te-value: "
            ,$hsp->P,"\n";

            push (@hspqstart, $hsp->query->start);
            push (@hspqend, $hsp->query->end);
            push (@hsphstart, $hsp->hit->start);
            push (@hsphend, $hsp->hit->end);
            @hspqstart = sort {$a <=> $b} @hspqstart;
```

```
                @hspqend = sort {$a <=> $b} @hspqend;
                @hsphstart = sort {$a <=> $b} @hsphstart;
                @hsphend = sort {$a <=> $b} @hsphend;
                 }

            # Print starting and ending values for query and hit:
                print "\nWorking protein start:\t@hspqstart\n";
                print "Working protein end:\t@hspqend\n\n";
                print "hit start:\t@hsphstart\n";
                print "hit end:\t@hsphend\n\n";

            # Remove the first and last numbers from each array
            # (representing beginning and end of all alignments).
                my $qstart = shift @hspqstart;
                my $qend = pop @hspqend;
                my $hstart = shift @hsphstart;
                my $hend = pop @hsphend;
                print "$wpid represents $blastdir
                from\t$hstart\tto\t$hend\n";
                print "Select working protein sequence
                from\t$qstart\tto\t$qend\n\n";
                }
              }
            }
        $totalselfhits += $selfhitcount;
        $totalwpcount += $wpcount;
        print "Number of files with no hits found:\t$nohit\n";
        print "Number of self-hits for $blastdir:\t$selfhitcount\n";
        print "Total working proteins for $blastdir:\t$wpcount\n\n";
        }
  print "Total number of working proteins:\t$totalwpcount\n";
print "Total working proteins finding
target:\t$totalselfhits\n";
}
```

```
# Place id numbers for successful working proteins in a hash, sort
# and print out id numbers. Do the same for target accession
# numbers.

print "Successful working proteins:\n";
foreach my $newwpid (sort keys %wpsuccess) {print " $newwpid"};
print "\nSuccessful targets:\n";
foreach my $newid (sort keys %tgtsuccess) { print " $newid"; };
print "\n";
foreach my $genwpid (sort keys %wpidhash) {print " $genwpid"};
print "\nWorking proteins processed:\n";
```

Section C.2 Sample BLAST parser output. *In-silico* proteins have an alphanumeric identifier, i.e. 3NT_005543R, and are called "working proteins" in this output.

```
Hit: NP_000010 acetyl-Coenzyme A acetyltransferase 1 precursor;
acetoacetyl Coenzyme A thiolase [Hs] - Hs, 427 aa.
Hit accession no:       NP_000010
Working protein created from: NP_000010
HSP: 381    437   %ID: 91.2   score: 270  bits: 127   e-value: 0.0
HSP: 599    653   %ID: 98.1   score: 270  bits: 127   e-value: 0.0
HSP: 320    368   %ID: 100    score: 255  bits: 120   e-value: 0.0
HSP: 661    699   %ID: 100    score: 199  bits: 94.9  e-value: 0.0
HSP: 148    188   %ID: 97.5   score: 188  bits: 89.9  e-value: 0.0
HSP: 253    290   %ID: 92.1   score: 183  bits: 87.5  e-value: 0.0
HSP: 472    521   %ID: 72     score: 178  bits: 85.2  e-value: 0.0
HSP: 204    236   %ID: 100    score: 169  bits: 81.1  e-value: 0.0
HSP: 520    563   %ID: 63.6   score: 125  bits: 60.7  e-value: 8e-15
HSP: 561    583   %ID: 100    score: 114  bits: 55.6  e-value: 0.0
HSP: 31     54    %ID: 100    score: 108  bits: 52.8  e-value: 0.0
HSP: 84     111   %ID: 67.8   score: 86   bits: 42.7  e-value: 8e-15

Working protein start:  31 84 148 204 253 320 381 472 520 561 599 661
Working protein end:    54 111 188 236 290 368 437 521 563 583 653 699

hit start:  1 24 40 80 108 145 192 240 275 314 336 389
hit end:    24 51 80 112 145 193 248 289 318 336 390 427

11NT_009151F represents NP_000010 from    1     to    427
Select working protein sequence from      31    to    699
```

```
Hit: NP_000010 acetyl-Coenzyme A acetyltransferase 1 precursor;
acetoacetyl Coenzyme A thiolase [Hs] - Hs, 427 aa.
Hit accession no:       NP_000010
Working protein created from: NP_000010
HSP: 9      58    %ID: 30     score: 60   bits: 30.1  e-value: 0.32

Working protein start:  9
Working protein end:    58

hit start:  174
hit end:    223

15NT_024749R represents NP_000010 from    174   to    223
Select working protein sequence from      9     to    58
:     :     :     :     :     :     :     :     :     :     :     :
:     :     :     :     :     :     :     :     :     :     :     :
:     :     :     :     :     :     :     :     :     :     :     :
Number of files with no hits found: 1
Number of self-hits for NP_000010:  7
Total working proteins for NP_000010:     35

Hit: NP_000013 adenosine deaminase; adenosine aminohydrolase [Hs] - Hs,
363 aa.
Hit accession no:       NP_000013
Working protein created from: NP_000013
HSP: 125    172   %ID: 100    score: 252  bits: 119   e-value: 0.0
HSP: 631    674   %ID: 100    score: 230  bits: 108   e-value: 0.0
HSP: 47     97    %ID: 86.2   score: 225  bits: 106   e-value: 0.0
HSP: 262    304   %ID: 95.3   score: 217  bits: 102   e-value: 0.0
HSP: 203    241   %ID: 97.4   score: 203  bits: 96.5  e-value: 0.0
HSP: 445    480   %ID: 94.4   score: 186  bits: 88.7  e-value: 0.0
HSP: 713    750   %ID: 94.7   score: 179  bits: 85.4  e-value: 0.0
HSP: 584    607   %ID: 95.8   score: 137  bits: 66.1  e-value: 0.0
HSP: 348    373   %ID: 96.1   score: 126  bits: 61.0  e-value: 0.0
HSP: 6      27    %ID: 95.4   score: 113  bits: 55.0  e-value: 0.0

Working protein start:  6 47 125 203 262 348 445 584 631 713
Working protein end:    27 97 172 241 304 373 480 607 674 750

hit start:  11 26 74 122 160 201 226 259 282 326
hit end:    32 76 121 160 202 226 261 282 325 363

20NT_011362R represents NP_000013 from    11    to    363
Select working protein sequence from      6     to    750

Number of files with no hits found: 0
Number of self-hits for NP_000013:  1
Total working proteins for NP_000013:     3

Total number of working proteins:   38
Total working proteins finding target:    8
```

REFERENCES

Altschul , S. F. et. al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 25, 3389-3402.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., et. al. (2002). The Pfam protein families database. Nucleic Acids Research 30, 276-280.

Branca, M. (2001). Streamlining drug discovery with breakthrough technologies for genomic target identification and validation. Retrieved October 1, 2002 from Cambridge Healthtech Institute: <http://www.chireports.com/content/articles/targetart.asp>.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. (2001). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge, U.K.: Cambridge University Press.

Eddy, S. R. (2001). HMMER: Profile hidden Markov models for biological sequence analysis. Retrieved October 30, 2002. <http://hmmer.wustl.edu>.

Federsel, H. (2001). Too many targets, not enough target validation. Drug Discovery Today, 6, 397-398.

Hardman, J. G., Limbird, L. E. & Goodman Gilman, A. (Ed.). (2002). *Goodman & Gilman's the pharmacological basis of therapeutics.* New York: McGraw-Hill.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature, 409, 860-921.

Kent, W.J. & Haussler, D. (2001). Assembly of the working draft of the human genome with GigAssembler. Genome Research 10.1101/gr.183201 <http://www.genome.org/cgi/doi/10.1101/gr.183201>.

Lewin, B. (1994). *Genes V.* Oxford, U.K.: Oxford University Press.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. Science, 227, 1435-41.

Lipman, D. J. & Pearson, W. R. (1988). Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences, 85, 2444-2448.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods of Enzymology, 183, 63-98.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. Journal of Molecular Biology, 276, 71-84.

Pearson, W. R. (1998). Flexible sequence similarity searching with the FASTA3 program package. Retrieved October 15, 2002. <http://www.people.virginia.edu/~wrp/pearson.html>.

Pevzner, P. A. (2001). Assembling puzzles from preassembled blocks. Retrieved October 30, 2002. Genome Research, 1461-1462. <http://www.genome.org/cgi/doe/10.1101/gr.206301>.

Sanseau, P. (2001). Impact of human genome sequencing for *in-silico* target discovery. Drug Discovery Today, 6, 316-323.

Schwager, C. (2002). 6-frame DNA to protein translation. [jpg image]. Retrieved November 1, 2002. <http://pc-ansorge11.embl-heidelberg.de/geneSkipper>.

Sterky, F. & Lundeberg J. (2000). Sequence analysis of genes and genomes. Journal of Biotechnology, 76, 1-31.

Swindells, M. B. & Overington, J. P. (2002). Prioritizing the proteome: Identifying pharmaceutically relevant targets. Drug Discovery Today, 7, 516-521.

Wilbur, W. J. & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. Proceedings of the National Academy of Sciences, 80, 726-730.

Wood, T.C. & Pearson, W. R. (1999). Evolution of protein sequences and structures. Journal of Molecular Biology, 291, 977-995.

Bioperl 1.0.2 [Perl-based software]. (2002, July 15). <http://bioperl.org>.

DrugBank, University of Alberta. Retrieved August 20, 2002.
<http://www.drugbank.ca>.

Investigational Drugs Database (release 2). Retrieved August 15, 2002.
<http://www.current-drugs.com>.

UCSC Genome Bioinformatics. (2002). University of California, Santa Cruz.
Retrieved September 30, 2002.
<http://www.genome.ucsc.edu>.

Swissprot (release 40.0). ExPASy (Expert Protein Analysis System). Swiss
Institute of Bioinformatics Retrieved July 1, 2002.
<http://www.expasy.org>.

Refseq. National Center for Biotechnology Information. Retrieved July 1, 2002.
<http://www.ncbi.nlm.nih.gov>.

RepeatMasker documentation. Retrieved November 25, 2002
<http://ftp.genome.washington.edu/RM/RepeatMasker.html>.