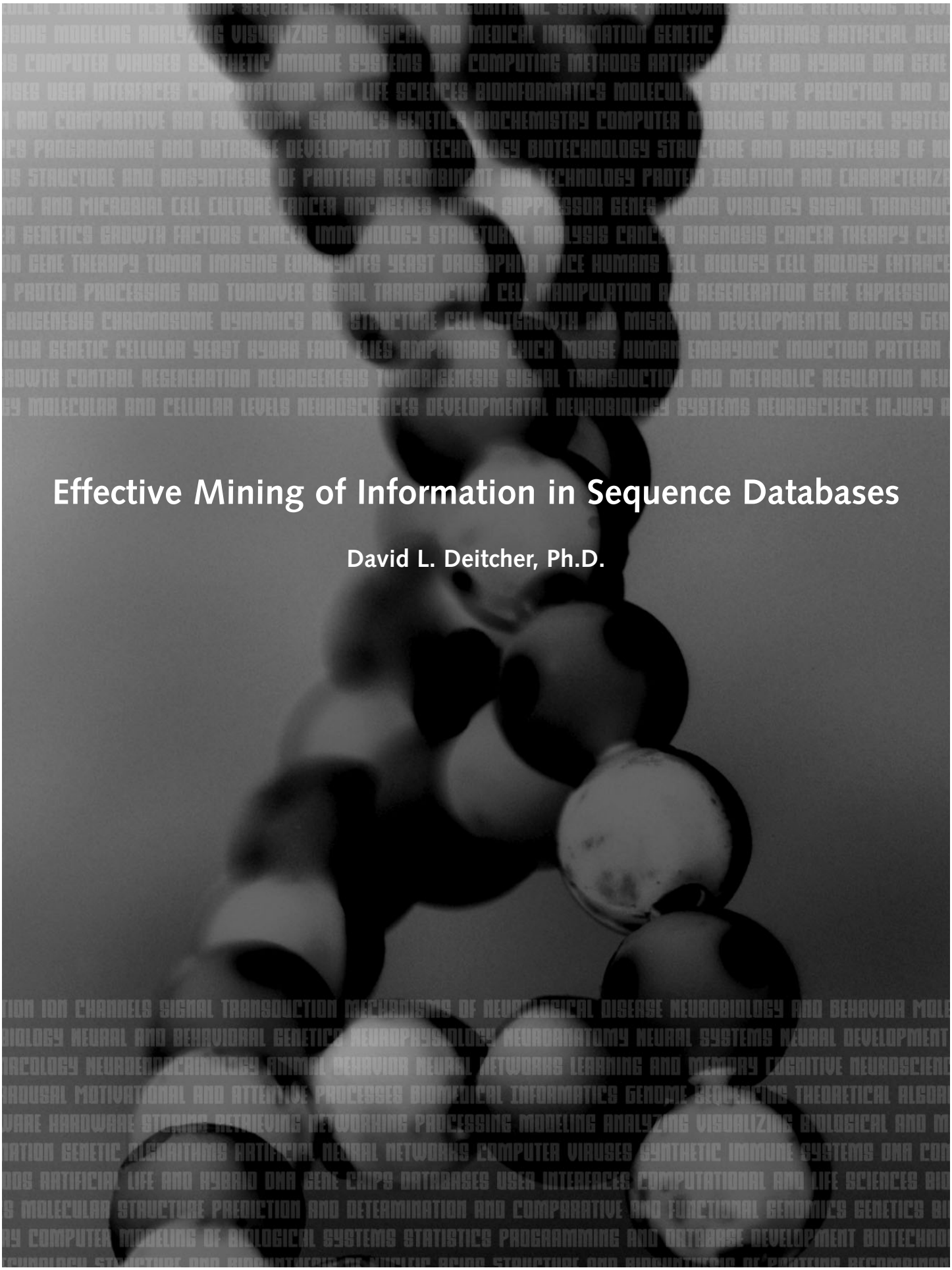

Effective Mining of Information in Sequence Databases

David L. Deitcher, Ph.D.

I. INTRODUCTION

The sequencing of the genomes of *Drosophila, C. elegans*, mouse, and humans has opened a floodgate of information, much of which is very useful to neurobiologists. The remarkable conservation of mechanisms governing the function of the nervous system has made the study of diverse experimental organisms extremely relevant to understanding how the human brain works. Whether you are working with one of the model organisms that have been sequenced or nonstandard organisms, genome sequence information can be a great asset to your research. However, understanding how the information is gathered, assembled, and analyzed is crucial in deciding how to proceed. Historically, many of the genes cloned were characterized in publications in which the authors determined the initiation ATG, transcript size, expression pattern, and the gene's biological role. Much of the current data on newly sequenced genes lacks this useful contextual information. The two principle types of unpublished sequence data are genomic sequences and expressed sequence tags (ESTs).

II. GETTING STARTED

Before beginning to explore the newly sequenced and uncharacterized genes from the database, it is very useful to learn as much as possible from the conventional scientific literature. Often, related genes from another organism have been examined or related genes from your organism have already been studied. Use of Entrez-PubMed or similar web-based literature searching will save a huge amount of time in the long run.

Why spend time looking in conventional literature when sequences can be acquired online in no time? The information gathered by large scale genomic and EST sequencing efforts has not been sufficiently examined by scientists, but rather, computers have assembled the information. This information has the appearance of being as solid and trustworthy as published work but very little of it has actually been checked for inconsistencies. To circumvent this issue, it is very important to determine how the information you are using has been gathered and whether it makes sense with other biological information. Thus, a good understanding of the biology of your system and the genes that may be involved in it will enable you to distinguish between a scientific time sink and a breakthrough.

Scientific papers give useful information such as expression pattern (both spatial and temporal), transcript size, alternative splicing, transcriptional start and stop sites, predicted motifs, chromosomal location, and homology to other proteins or genes. Since the authors have usually thought about their gene of interest, connections to other systems and genes are often discussed.

III. GENOMIC SEQUENCES

The sequencing of the whole genomes has utilized two related approaches. In one approach, large genomic clones such as BAC clones are sequenced and mapped to chromosomal locations. Eventually overlapping clones are used to assemble the full sequence. An alternative approach is to generate random fragments of genomic DNA, sequence the many random clones, and assemble the overlapping sequences into contigs. For example, *Drosophila* has mainly been sequenced by the random clone approach with an average of 10-12 fold coverage of the nonrepetitive DNA sequences of the genome (1). This highly redundant sequencing approach results in a very low error rate.

Are the whole human or *Drosophila* genomes fully sequenced? No. Some DNAs are likely to be missed in sequencing efforts since they were underrepresented in either the BAC clone collection or fragmentation approach due to efficiency of subcloning or unstable sequences. Additionally, DNA containing very repetitive DNA (like heterochromatin) is difficult to sequence and assemble. In *Drosophila*, very little of the heterochromatic regions of the genome have been sequenced though efforts in this direction have begun. These regions are not devoid of genes encoding proteins. For example, the *Drosophila SNAP-25* gene resides in heterochromatin on the left arm of the third chromosome (2). My laboratory and others have isolated lethal complementation groups from heterochromatic DNA. Such genes are often placed in incorrect chromosomal locations due to the highly repetitive DNA that surrounds them. Such is the case with *SNAP-25*, where initial *in situ* hybridization placed the gene in an entirely different chromosomal location. In addition to the heterochromatin issue, very similar but distinct DNA sequences may be assembled together as if they are the same sequence. When such sequences are assembled, stretches of unique DNA may be omitted since they appear to be identical to overlapping clones. While these are significant problems, for the most part, the large scale sequencing efforts produce highly accurate sequence.

IV. GENE PREDICTION

Once the sequence of a genome is assembled it is then possible to predict where the genes reside. In prokaryotes, the gene prediction programs work better due to a lack of introns and less variability in cis control elements. In eukaryotes, this situation is more complex. Intron/exon boundaries are difficult to predict, exons may be alternatively spliced, and cis control elements sequences may act distantly from the gene. Even the TATA box is present in only about 70% of promoter sequences (3). Programs which predict genes rely on several complementary methods but such programs often give differing predictions. The accuracy of the prediction is often related to whether the features of the gene match previously identified genes. Thus, they may not be a good predictor of unknown genes. Caution should be used in using these tools without other confirmatory information.

V. EST CLONES

The deficit of gene predicting programs has been partially filled by EST clones. If a predicted gene has a matching EST clone then it is more likely to be an actual gene. EST clones are cDNA clones that are sequenced from the *5'* or *3'* end. Generally, the sequence is done once so the error rate is approximately 1 in 100, far higher than redundant genomic sequencing (4). EST clones have the same pitfalls as all cDNA clones. Many enzymatic steps are involved in generating a cDNA library, some of which may not go to completion, resulting in truncated cDNAs or cloning artifacts (Figure 1). It is a good idea to obtain more than one EST clone that matches your gene of interest. If you sequence two or more independent clones and they give the same sequence then it is likely that a given EST clone is not aberrant. Another problem with EST clones it that they may be chimeric (Figure 1, right). This problem can also be addressed by analyzing more than one independent EST clones.

ESTs offer several other advantages over gene prediction. The source of RNA is indicated for EST clones so one can at least know that what tissue expresses the gene. EST clones are also extremely useful in that instead of having to laboriously screen a cDNA library; one can simply order the clone. Also, by sequencing a number of ESTs one can also assess whether alternative splicing is occurring in your gene of interest.

VI. GENE PREDICTION, EST CLONES, AND SHAKER

Shaker was the first K+ channel to be cloned (5). It is probably one of the best studied genes in *Drosophila* and it was selected for discussion because much was already known about the gene prior to the genome sequencing effort. The *Shaker* locus spans a very large region on the X chromosome. The *Shaker* gene produces transcripts that have alternative 5'ends and alternative splicing at the 3' end resulting in proteins with different N- and C termini (6). In Gadfly, the Celera site for the *Drosophila* genome, one can access information about *Shaker* by looking at polytene chromosome band 16, the map position of *Shaker*. On the *Shaker* gene web page, the predicted gene is shown, its identification as a voltage-gated K+ channel is listed, and several EST clones that match the *Shaker* sequence are also presented. If one examines the gene prediction of Shaker, no mention is found of alternative splicing, though this is well-established. A BLAST search of the 5' EST sequence of clone GH15217 (one of the *Shaker* EST clones) reveals that the number 1 match is indeed *Shaker*. However, if the sequence from the 3' end of GH15217 is used to perform a BLAST search, it matches a methionyl aminopeptidase. Evidently, EST clone GH15217 is a chimeric clone, part *Shaker* part encoding methionyl aminopeptidase. A careful search of the Berkeley Drosophila Genome Project site warns that two of the libraries used for the EST collection contain approximately 25% chimeric clones. Newer cDNA libraries are thought to have a lower incidence of chimeric clones. Before ordering an EST clone it is useful to run your own BLAST search and make sure the 5 and 3' ends of the clone match the same gene. It is also prudent to order several EST clones and fully sequence them before using them for your research. The errors presented here are largely because most of the data has been subject to very limited human curation. I expect that other genome sequences have similar problems. Annotation corrections are possible at the website so later genome release versions should have many fewer errors.

VII. USING BIOINFORMATICS TO CLONE GENES

The procedure for identifying new genes depends on the particulars of your experimental system. If you want to identify a new member of a gene family this could be done by a simple BLAST search. First, locate the known family member genes or proteins using Entrez-Nucleotide or Entrez-Protein at the NCBI website **http://www.ncbi.nlm.nih.gov/**. To limit the number of entries, it is useful to include the species. For finding DNA sequences, the terms "cds" (coding sequence) or "mRNA" will limit the number of patents and unrelated entries that are retrieved. Once an entry has been identified, simply copy the translation sequence and perform a TBLASTN search to identify whatever nucleotide sequences matches the query. If you are lucky, you may find multiple EST clones of a new gene family member.

If however no EST clones have been identified for your gene of interest, then one can use other gene family sequences and RT-PCR to clone the gene. I will give two examples of using this approach. In the first case, I was involved in cloning a new NMDA receptor subunit gene from rat. At the time, NMDAR1 and NMDA2A-2D had been identified (7, 8). Sequence homologies are often concentrated in local regions so we attempted to identify two different regions of the NMDA receptor subunit gene family that were highly conserved. Since we planned to use PCR to find this additional family member, conserved amino acid sequences that were devoid or low in the number of Arg's, Leu's, or Ser's were selected for primer design. One conserved stretch was selected based upon the finding that this region contained two conserved cysteines that determine redox modulation of the NMDA receptor (9) (Figure 2). The primers selected would amplify all previously described NMDA receptor subunits. RNA was prepared from several different neural tissues, it was reverse transcribed and amplified by PCR. The resulting product was subcloned sequenced. One clone was found to contain a novel NMDA receptor subunit. This clone was then used as a probe to screen a cDNA library. A new NMDA receptor subunit was identified called NMDAR-L (10), which has been renamed NR3A. The amino acid identity of NR3A with other NMDA subunits was only approximately 27%. Thus, selecting small, but highly conserved regions was critical for its identification.

How might one go about cloning a gene from a nonstandard organism? One method is to use homologous genes from the evolutionarily-related species to design PCR primers. In this project we attempted to identify the Midshipman aromatase gene, a gene thought to be involved in the sexual and behavioral dimorphisms among male subtypes of the species. We used aromatase sequence data from two more widely studied fish, the goldfish and tilapia. These two sequences were used in a BLAST of two sequences against each other to identify areas of homology (Figure 3). Several regions were selected with homologous sequences and PCR primers were designed. Many of the primer sets failed to amplify anything; however, one primer set near the C terminus gave the expected size band. The product was subcloned and sequenced and a Midshipman specific clone was identified. Based upon the DNA sequence, *in situ* hybridization was performed and a Midshipman-specific antibody was used to localize aromatase mRNA and protein respectively. Contrary to expectations, aromatase was specifically localized to glial cells, not neurons (11).

VIII. SEQUENCES FOR ANTIBODY PRODUCTION

Cloning of genes facilitates the study of the proteins that are encoded by those genes. One of the most important uses of sequence information is to generate an immunogen for antibody production. Antibodies can be made to fusion proteins or to synthetic peptides. Prior to the sequencing of whole genomes, the selected immunogen would often have sequence homology to other proteins and this would result in undesirable protein cross-reactivity. Now it is possible to design an immunogen with almost no chance of cross-reaction within a given species.

Bioinformatics can also be used to predict whether an antibody raised against a related protein will crossreact with your protein. By comparing the immunogen sequence from academic papers and company literature to your protein sequence, one may be able to find antibodies raised to related proteins from other species that could be useful in your research. This can greatly expand the number of antibodies available.

Sometimes it is useful to measure the relative amounts of two related proteins. This is very difficult if different antibodies are employed as they have different binding affinities. In some cases, it may be possible to find a large enough protein sequence that is 100% conserved between two proteins and use this as an immunogen for antibody production. For example, in my laboratory, we designed such an antibody that equally recognizes the related proteins SNAP-24 (12) and SNAP-25. Thus, with such an antibody, the relative amounts of each protein can be determined by western blot (if they have different molecular weights). A similar approach could be employed in studying multisubunit proteins, such as ion channels. A determination of the relative amounts of each type of subunit could be valuable tool for determining channel composition.

IX. MOTIFS

Sequences may also be analyzed for protein motifs. Prosite is a commonly used program for identification of protein sequence motifs. Two other related domain searching sites are found on the main NCBI BLAST web page, RPS-BLAST and DART. These programs focus more on showing families of related proteins, based upon conserved features. Many of the programs with motif searching require the protein sequence to be in FASTA format. Any cut and pasted sequence can easily be converted to FASTA at the Baylor College of Medicine Sequence Utilities site (listed below). In order to understand your motif results consult **http://smart.embl-heidelberg.de/smart/domain_table.cgi** which is one of the most comprehensive listing of protein motifs. Investigating the function of motifs found in your protein may provide useful hints as to what proteins interact with each other and what their function is.

X. USEFUL WEBSITES

http://www.ncbi.nlm.nih.gov/
Home page of NCBI from which most homology searching programs can be found

http://dot.imgen.bcm.tmc.edu:9331/seq-util/seq-util.html
Baylor Sequence Utilities website – useful for FASTA conversion, translations, and restriction mapping

http://www.ncbi.nlm.nih.gov:80/BLAST/
General BLAST page (select appropriate type of BLAST search)

http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html
BLAST of two sequences for highlighting areas of identity

http://hits.isb-sib.ch/cgi-bin/PFSCAN
Prosite scan for identifying protein motifs

http://smart.embl-heidelberg.de/smart/domain_table.cgi
Motif index - extensive index of protein motifs

http://www.ch.embnet.org/software/ClustalW.html
Multiple alignments (requires FASTA format)

http://www.expasy.ch/tools/pi_tool.html
Calculation of pI and MW of proteins

http://alces.med.umn.edu/rawtm.html
Tm of primers for PCR

http://www.fruitfly.org/
Berkeley Drosophila Genome Project Homepage

XI. REFERENCES

1. M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne et. al. Science 287, 2185 (2000).
2. S.S. Rao et al., EMBO J. 20, 6761 (2001).
3. A.D. Baxevanis, "Predictive Methods Using DNA Sequences" in Bioinformatics, Wiley and Son, 2001.
4. D.W. Mount, Bioinformatics, Cold Spring Harbor Laboratory Press, 2001.
5. D.M. Papazian, T.L. Schwarz, B.L. Tempel, Y.N. Jan, L.Y. Jan, Science, 237,749 (1987).
6. T.L. Schwarz, B.L. Tempel, D.M. Papazian, Y.N. Jan, L.Y. Jan, Nature, 331, 137 (1988).
7. K. Moriyoshi et. al., Nature, 354, 31 (1991).
8. H. Monyer et al., Science , 256, 1217 (1992).
9. J.M. Sullivan et al., Neuron 13, 929 (1994).
10. N.J. Sucher et al, J. Neurosci., 15, 6509, (1995).
11. P.M. Forlano, D.L. Deitcher, D.A. Myers, A.H. Bass. J. Neurosci. 15, 8943 (2001).
12. B.A. Niemeyer, T.L. Schwarz, J. Cell. Sci. 113, 4055 (2000).
13. J.S. Nelson, Fishes of the World, Wiley and Sons, 1994.