

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

University of Alberta

**Bioinformatics in Gene Finding and Database Design: A
Pharmaceutical Approach**

by

Haiyan Zhang 

A thesis submitted to the Faculty of Graduate Studies and Research in partial
Fulfillment of the requirements for the degree of

Master of Science

In

Pharmaceutical Sciences

**Faculty of
Pharmacy and Pharmaceutical Sciences
Edmonton, Alberta
Spring, 2002**



**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**385 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**385, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-69784-3

Canada

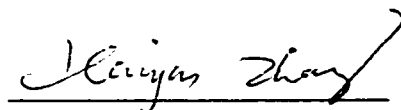
University of Alberta

Library Release Form

Name of Author: Haiyan Zhang
Title of Thesis: Bioinformatics in Gene Finding and Database Design
-- Applications in Pharmaceutical Research
Degree: Master of Science
Year this Degree Granted: 2001

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis or any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



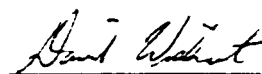
Faculty of Pharmacy and
Pharmaceutical Sciences
University of Alberta
Edmonton, Alberta
T6G 2N8


Date: Jan. 22, 2002

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Bioinformatics in Gene Finding and Database Design: A Pharmaceutical Approach" by Haiyan Zhang in partial fulfillment of the requirements for the degree of Master of Science in Pharmaceutical Sciences.


Dr. David S. Wishart


Dr. Mavanur R. Suresh


Dr. Russell Greiner

Date: January 22, 2002

Abstract

Over the past 15 years, bioinformatics has played an increasingly important role in pharmaceutical research. This dissertation presents two bioinformatics applications that may have particular relevance to drug discovery or development. The first application (Chapter 2) describes the development of two new algorithms (called GRPL and GRPL+) to improve the accuracy of gene prediction in eukaryotic DNA. Specifically, this set of computer program combines reference point logistic (RPL) methods with “smart” sequence alignment methods to provide substantially improved accuracy in gene identification. As part of the evaluation of the GRPL program, we show how the quality of gene predictions can be improved with increasing database size. This technique has important implications for identifying known and unknown disease genes in the draft human genome sequence. Chapter 3 describes the development of self-updating, self-correcting databases containing biological or chemical information. Self-updating databases use data-mining and data-validation methods to automatically extract and deposit “corrected” electronic information into a database or archive. With the rapid increase in both the size and number of biological, chemical and pharmaceutical databases, there is a growing need to create automated methods to consolidate, update, validate and correct the data in these databases. In this thesis, I illustrate the development of such a self-updating, self-correcting database with a specific example called RefDB. RefDB is a database used by NMR spectroscopists to archive, access and analyze NMR chemical shift data.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. David Wishart, for his encouragement, enthusiasm and help. Without his mentoring, I would never have entered the field of bioinformatics. His amiable personality, his broad range of knowledge and acute insight into problems has been of indispensable assistance to my research.

I would also wish to express my whole-hearted thanks to all my colleagues and fellow students I have had the pleasure of associating with. In particular, I would like to thank Dr. Alexander Nip for his help at the beginning of my studies in computer programming. I would also like to thank Ryan de Los Angeles and Ashenafi Abera for their help in building the RefDB database.

I particularly wish to express my deep gratitude to my entire family, for their love, sacrifice, and encouragement over my graduate studies.

Finally, I would like to acknowledge BioTools Inc. (Edmonton), the Protein Engineering Network of Centres of Excellence (PENCE) and the Alberta Cancer Board for their financial and technical support throughout my graduate career.

Table of contents

	Page
Chapter 1: Introduction	1
1.1 Bioinformatics	1
1.2 Bioinformatics in Pharmaceutical Research	7
1.3 Outline of Dissertation	9
1.4 Protein Sequence Searching	10
1.5 Sequence Analysis – Exon Prediction	20
1.6 Bioinformatics Applications in NMR	23
1.7 My Contribution to This Dissertation	28
Chapter 2: Prediction of Genetic Structure in Eukaryotic DNA Using Reference Point Logistic Regression and Sequence Alignment	28
2.1 Introduction	30
2.2 Methods	36
2.3 Results and Discussion	40
2.4 Conclusion	50
2.5 Availability	52
Chapter 3: RefDB - A Database of Uniformly Referenced Protein Chemical Shift Assignments Derived from the BioMagResBank	54
3.1 Introduction.....	54
3.2 Materials and Methods.....	56
3.3 Results and Discussion.....	62
3.4 Conclusion.....	83
3.5 Availability	85
Chapter 4: General Discussion and Conclusion	86
Appendix A: Simpred	91
Appendix B: ShiftoR	97
References	106

List of Tables

	Page
Table 2.1	42
Table 2.2	42
Table 2.3	43
Table 2.4	44
Table 2.5	45
Table 2.6	47
Table 2.7	48
Table 3.1	67
Table 3.2	71
Table 3.3	76
Table 3.4	77
Table 3.5	78
Table 3.6	79
Table 3.7	80
Table 3.8	81
Table B.1	100
Table B.2	104

List of Figures

	Page
Figure 1.1	2
Figure 1.2	16
Figure 1.3	18
Figure 1.4	21
Figure 2.1	32
Figure 2.2	35
Figure 2.3	49
Figure 2.4	51
Figure 3.1	60
Figure 3.2	61
Figure 3.3	63
Figure 3.4	69
Figure 3.5	70
Figure A.1	93
Figure A.2	96
Figure B.1	99
Figure B.2	102
Figure B.3	103

List of Abbreviations

Amino Acids:

A(Ala)	=	L-alanine
C(Cys)	=	L-cysteine
D(Asp)	=	L-aspartic acid
E(Glu)	=	L-glutamic acid
F(Phe)	=	L-phenylalanine
G(Gly)	=	L-glycine
H(His)	=	L-histidine
I(Ile)	=	L-isoleucine
K(Lys)	=	L-lysine
L(Leu)	=	L-leucine
M(Met)	=	L-methionine
N(Asn)	=	L-asparagine
P(Pro)	=	L-proline
Q(Gln)	=	L-glutamine
R(Arg)	=	L-arginine
S(Ser)	=	L-serine
T(Thr)	=	L-threonine
V(Val)	=	L-valine
W(Trp)	=	L-tryptophan
Y(Tyr)	=	L-tyrosine

2D	Two Dimensional
3D	Three Dimensional
BLAST	Basic Local Alignment Search Tool
BMRB	BioMagResBank
CDS	Coding Sequence
CGI	Common Gateway Interface
CSI	Chemical shift index
DSS	2, 2-dimethyl-2-silapentane-5-sufonic acid
EST	Expressed segment tag
GPCR	G-protein coupled receptor
HSP	High scoring segment

HTML	HyperText Markup Language
IUB	International union of biochemists
IUPAC	The International Union of Pure and Applied Chemistry
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser effect
PDB	Protein databank
PIR	Protein international resources
ppm	Parts per million
TFE	Trifluoroethanol
TMS	Tetramethysilane
VADAR	Volume Area Dihedral Angle Reporter

Chapter 1

Introduction

1.1 Bioinformatics

What is Bioinformatics?

The field of *Bioinformatics* first emerged in the 1980s in response to the growing demand for computational techniques and computer resources to handle the explosion of molecular sequence data (Figure 1.1). Because of its growing importance, the term “*bioinformatics*” has been commandeered by several different disciplines to mean rather different things. Strictly speaking, it is a field of information technology that endeavors to improve the storage, management and analysis of biological information. Practically speaking, it is concerned with the handling and analysis of DNA and protein sequence data. Although the term bioinformatics was first coined in the 1980s (http://www.d-trends.com/webs/bio_1.html), the idea of using computers to store and manage biological data was actually initiated by X-ray protein crystallographers in the 1960’s (Levinthal, 1966). Their early work led to the establishment of the first bioinformatics database in 1971 - Brookhaven National Laboratory’s Protein DataBank (PDB), a database of 3D protein structures (Bernstein et al., 1977).

However, the advent of what we call bioinformatics today was mainly driven not by X-ray crystallography but by the development of improved DNA sequencing technology (Maxam and Gilbert, 1977; Sanger et al., 1977). Prior to these Nobel-prize

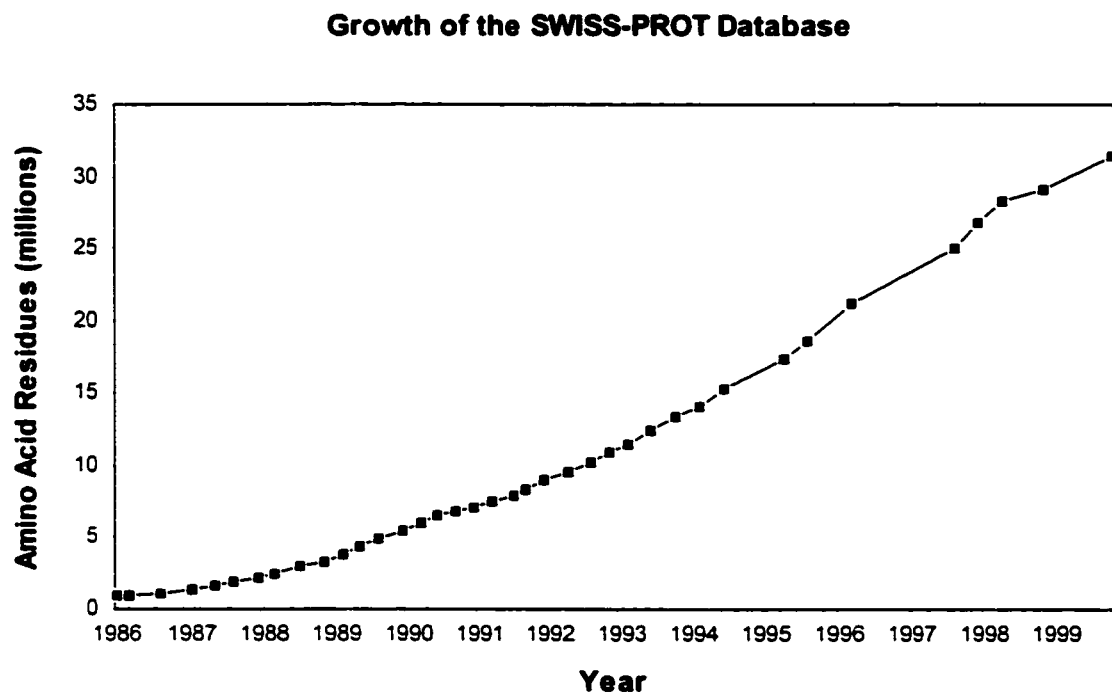


Figure 1.1 The exponential growth in the number of the total amino acid residues deposited in the SWISS-PROT protein database.

winning developments, it would take a laboratory at least two months to sequence just 150 nucleotides (Moukheiber, 1998). By the late 1970s – courtesy of Gilbert and Sanger - it was possible to sequence ~200 bases per day. With the development of fluorescence labeling technology (Wilson et al., 1990) and the introduction of multiplexed capillary electrophoresis (Cohen et al., 1990; Luckey et al., 1990; Swerdlow and Gesteland, 1990), fully automated DNA sequencers soon appeared. Now with instruments such as the ABI 3700 or the Pharmacia Megabace 500, it is possible to sequence 500,000 bases per day on a single machine! Today, companies such as Celera, Incyte, Monsanto and others are capable of sequencing up to 100 million bases a day.

Because of this new technology, DNA sequencing activities became heavily dependent on computer software for assembling, storing, and managing DNA sequence data. The rapid accumulation of DNA sequence data also stimulated much interest in the development of statistical methods and computer programs for analyzing DNA and protein sequence data. The need for computational tools was especially amplified with the launch of the Human Genome Project in 1990 (Deloukas et al., 1998). Beginning as a 15-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health, its ultimate goal is to map, sequence and identify all 30,000+ genes in the human genome. The first draft of human genome was completed on June 25, 2000 (<http://www.ornl.gov/hgmis/project/clinton1.html>) and released publicly on Feb 15, 2001 (McPherson et al., 2001; Venter et al., 2001). It is expected that a “finished” version of human genome will be released in 2003. By then it is also expected that the genomes of more than 500 other organisms will have been sequenced. Not only has bioinformatics

played a key role in handling, sorting and storing this genomic information, it is also expected to help with the new challenges ahead in inferring gene and protein functionality. It is this latter application which will be critical in identifying potential medical applications or pharmaceutical targets.

The second reason for the rapid emergence of the bioinformatics as a major force in biology can be attributed to the spectacular growth in computing technology. As uncannily predicted by Gordon Moore in 1965: "The processing speed of a microchip will double about every 18 months". Today, this trend still holds true and it is known as Moore's Law. Such a rapid rate of computer hardware development has led to the creation of a thriving computer industry that delivers very high performance machines that are relatively inexpensive. This, in turn, has led to the ubiquitous distribution of desktop computers, allowing easy access to computational tools among biologists and genome researchers.

Another very significant reason for the rapid growth in computer use among genome researchers has been the appearance of the "Information Superhighway" (i.e. the Internet). Originally developed in 1969 by the U.S. Department of Defense for research into communication networking, ARPANET (as it was called then) grew from text-only messaging system to a graphics-rich, interactive communication medium, enabling rapid information exchange. By 1993, Internet uses exploded with the introduction of browsers such as Mosaic and Netscape. These web browsers and their special communication language called HTML greatly facilitated access and communication between

individuals, research labs, universities and other large research organizations. Dedicated bioinformatics web servers such as EXPASY (<http://www.expasy.ch>) (Appel et al., 1994) and the National Center for Biotechnology Information (NCBI) web site (<http://www.ncbi.nlm.nih.org>) (Jenuth, 2000) heralded the establishment of the Internet as the primary means of communication among genome researchers, causing the field of bioinformatics to truly take off.

Current Status of Bioinformatics

Bioinformatics is now being practiced worldwide by thousands of individual researchers, academic groups, companies, national and international research consortia. With the technology of DNA sequencing well in hand, and with nearly 100 genomes deposited in various databanks around the world, recent studies suggest that the most pressing tasks in bioinformatics now involve genome annotation and functional identification (Stevens et al., 2001). Consequently much of the bioinformatics research of today focuses on the following areas:

- Finding or identifying genes in the DNA sequences of prokaryotic and eukaryotic organisms. (e.g. GENSCAN (Burge and Karlin, 1997), GENMARK (Borodovsky, 1993), GRPL (Hooper et al., 2000))
- Developing methods to predict the structure and/or function of newly discovered proteins. (e.g. sequence alignment tools, such as PSI-BLAST (Altschul et al., 1997) and BLAST (Zhang and Madden, 1997))
- Clustering protein sequences into families of related sequences. (e.g. PROSITE (Bairoch, 1997), PFAM (Bateman et al., 2000), BLOCKS (Henikoff et al., 1999), DOMO (Gracy and Argos, 1998), PRINTS (Attwood et al., 1998))
- Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships. (e.g. PHYLIP (Retief, 2000))

- Predicting secondary and tertiary structures from protein sequences. (e.g. PHD (Rost, 1996))

Various bioinformatics tools have been developed to meet these requirements, and many of them are freely available over the Internet. In fact, according to the European Bioinformatics Institute's (EBI, Cambridge, UK), there are more than 600 of these tools (<http://www.embl-ebi.ac.uk/biocat>), and their number is growing rapidly.

Bioinformatics is now recognized as an essential tool for dealing with complete genome sequences, for extracting gene coding sequences, for identifying their corresponding protein sequences, for performing multi-sequence comparisons across species, and for predicting or modeling three-dimensional protein structures. It is believed, however, the real long-term value of bioinformatics lies not so much in the creation of analytical tools, but in the conversion of the knowledge that bioinformatics delivers into safer foods, better drugs, improved therapeutics, and ultimately an improved quality of life.

The Future of Bioinformatics

The ultimate goal of the Human Genome Project is to understand the functioning of living organisms at the molecular, cellular and supracellular level. Such an understanding holds enormous promise for the early detection and treatment of disease. As has already been mentioned, the first objective in the Human Genome Project was to sequence the DNA of humans as well as a variety of related organisms. Now that the first draft of human genome has been completed, we have entered the second phase – a phase

that some have called “the post-genomic era”. The focus of this second phase of genomics will be to identify each gene, to ascertain its function, to determine its structure and identify its interacting partners. As a result, a new field within bioinformatics, called functional bioinformatics, has emerged.

In this new post-genomic era, we can expect that DNA microarrays and proteomics will likely play an increasingly important role. Both of these high throughput techniques depend critically on basic bioinformatics to manage the resulting data. Interestingly, these same bioinformatics tools are being combined with results from clinical trials, statistics, and population genetics to help in drug discovery and drug testing. Another major application of bioinformatics will be the modeling of genetic circuits and metabolic networks (Persidis, 1999). There are a number of interesting computational developments in cellular and subcellular simulation (Normile, 1999; Akutsu et al., 2000), which promise, one day, to deliver accurate, *in silico* models of biological function.

1.2 Bioinformatics in Pharmaceutical Research

Target discovery constitutes one of the main components of today's early stage pharmaceutical research. The aim of target discovery is to identify and validate suitable drug targets (i.e. proteins or genes) for therapeutic intervention. However, only a small fraction of human proteins are actually targeted by today's drugs. Indeed, reports published in 1997 estimate that current therapies for most genetic diseases target fewer than 500 of the 30,000+ different proteins in the human genome (Drews and Ryser, 1997).

The majority of these targets are receptors such as G-protein-coupled receptors (GPCRs), which account for 45% of all targets; metabolic enzymes account for 28% of the remainder. Some target classes are, therefore, more 'successful' or exhibit better tractability in the drug discovery process. In addition to identifying proteins amenable to small molecule targeting, therapeutic proteins or protein drugs (such as cytokines, growth factors and monoclonal antibodies) must also be considered when identifying novel targets in the human genome. The total number of tractable targets remains difficult to establish given the uncertainty surrounding the total number of human genes. However, it has been estimated that the number of drug targets is probably 5,000-10,000 (Drews, 2000). This number is 10 – 20 times greater than the current repertoire of drug targets. Clearly the fruits of the Human Genome Project will undoubtedly change how and where we look for new drugs and how we assess drug targets.

The Human Genome Project is also creating a similar revolution in the way scientists think about drug discovery for infectious diseases. Now it is increasingly possible to identify particular genes responsible for (or involved in) many infectious diseases (caused by bacteria, viruses or parasites) that cannot be treated with conventional antibiotics. By assisting with the identification of these genes that are key to viral replication or bacterial metabolism, bioinformatics allows scientists to start thinking about how to design drugs, which can turn the responsible genes (or proteins) "on" or "off". By predicting something about the structure, activity or location of a particular protein, bioinformatics software allows pharmaceutical scientists to narrow the search for appropriate lead compounds. This has been particularly true for researchers trying to find

drugs to combat several important viral diseases, including AIDS, hepatitis C and viral meningitis. Since 1990, there have been several successful examples of drugs that have been rationally identified and designed with the help of bioinformatics, including HIV protease inhibitors (DesJarlais et al., 1990; Olson and Goodsell, 1998), hepatitis C inhibitors (Sintchak and Nimmesgern, 2000), therapeutics for viral meningitis (Buttery and Moxon, 2000) and antitumor agents such as capecitabine (Verweij, 1999). The success that bioinformatics has had in identifying important disease causing genes and pointing the way to new lead compounds for drug design has encouraged many pharmaceutical companies to invest heavily into the areas of bioinformatics, proteomics and genomics to assist with their drug discovery program (Persidis, 1998).

In the next 10 years it is expected that many more genes responsible for debilitating human diseases will be uncovered using the search engines and the powerful analytical techniques developed by bioinformatics software and database specialists. The fact that bioinformatics is proving so useful in the identification of potential drug targets (and even potential drugs) has led some to suggest that by the year 2005, more biology and drug discovery will be done “*in silico*” than “*in vivo*” and that many “wet” labs will actually become “digital” labs which will conduct their experiments on computers rather than on rats (Sanseau, 2001).

1.3 Outline of Dissertation

As can be seen by the brief review, bioinformatics is a field that is expanding rapidly to meet the growing demands of both genomics technologies and genomics

researchers. As such, it is almost impossible to conduct research in all areas of bioinformatics. In light of the limitation, I have chosen to focus this thesis on a relatively narrow topic: That is, the novel application of sequence comparisons and the development of novel databases. In particular, I will describe several innovative bioinformatics programs which make use of protein sequence alignment to produce: 1) more accurate and precise gene structure prediction; and 2) a set of bioinformatic databases and sequence alignment programs to facilitate NMR spectroscopy of peptides and proteins.

Specifically, the working hypothesis for chapter 2 is that the reference point logistic regression combined with protein sequence alignment, dynamic programming and general hidden markov models can be more accurate than other techniques in predicting eukaryotic gene structure. In chapter 3, the working hypothesis is that protein chemical shifts can be calculated with sufficient accuracy such that chemical shift referencing and/or assignment errors can be detected and corrected.

Since each chapter has its own detailed introduction, the following discussion will be limited to some basic background information not covered in subsequent chapters.

1.4 Protein Sequence Searching

Protein Sequence Databases

The first step in analyzing sequence information is to assemble it into central, shareable resource, i.e. a database. Databases are, effectively, electronic filing cabinets, -- a convenient and efficient method of storing vast amounts of information. Bioinformatics

databases are essential tools for comparative analysis and data-mining. Historically, protein sequence databases preceded the appearance of nucleotide sequence databases. In the early 1960s, Dayhoff and colleagues collected all known peptide and protein sequences and compiled them into *the Atlas of Protein Sequence and Structure* (Dayhoff, 1965). This early protein database eventually led to the establishment of the PIR database (Barker et al., 2001).

By the early 1980s, sequence information started to become so abundant that it was impossible to archive it in catalogs, in books or in journals. Realizing this, several laboratories saw that there might be advantages to harvesting and storing these sequences in central computer repositories. The first of these electronic databases was GenBank, a DNA sequence databank established in 1982 (Benson et al., 2000). There are now many different types of electronic bioinformatic databases, depending both on the nature of the information being stored (e.g., sequences, structures, 2D gel images, etc.) and on the manner of the data storage. Here we concerned only with the different types of biological data, rather than on particular storage or management mechanisms. In the context of protein sequence analysis, there are two types of databases: primary and composite databases. We shall discuss primary databases first. Primary sequence databases contain amino acid sequences, stored as linear character strings (using the IUPAC single letter mode) denoting the constituent residues. The major protein primary databases include the Protein Information Resource (PIR) (Barker et al., 2001), SWISS-PROT (Bairoch and Apweiler, 2000) and TrEMBL. A brief summary of each of these primary databases is given below:

The PIR Database

The Protein Information Resource (PIR) was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers in the identification and interpretation of protein sequence information. In its current form, the database is split into four distinct sections (designated PIR1, PIR2, PIR3 and PIR4) which differ in terms of the quality of data and level of annotation provided. Since 1988, the PIR database has been maintained collaboratively by PIR-International (Barker et al., 2001). This is an association of macromolecular sequence data collection centers including the Protein Information Resource (PIR) at the NBRF, the International Protein Information Database of Japan (JIPID), and the Martinsried Institute for Protein Sequences (MIPS).

The SWISS-PROT Database

In 1986, SWISS-PROT (Bairoch and Apweiler, 2000) was produced collaboratively by the Department of Medical Biochemistry at the University of Geneva and the European Molecular Biology Laboratory (EMBL). The SWISS-PROT database endeavors to provide high-level annotations, including descriptions of the function, 3D structure, domain structure, post-translational modifications, variants, and extensive references. SWISS-PROT aims to be minimally redundant and is interlinked to many other bioinformatics resources. Currently, the database is maintained collaboratively by the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI) and EMBL.

The TrEMBL Database

In 1996, a computer-annotated supplement to SWISS-PROT was created, termed Translated EMBL (TrEMBL) (Bairoch and Apweiler, 2000). This protein sequence database is in the same format as SWISS-PROT and contains translations of all coding sequences (CDS) contained in the EMBL gene sequence database (the European mirror of GenBank). TrEMBL was designed to address the need for a well-structured SWISS-PROT-like resource that would like to allow very rapid access to protein sequence data from genome projects, without having to compromise the quality of SWISS-PROT itself by incorporating sequences with insufficient analysis and annotation.

There are now more than 100 primary protein sequence databases available (Discala et al., 2000), so it is becoming increasing difficult to know which database to use and when. One solution to this problem is to compile a composite database, which amalgamates a variety of different primary sources. Different strategies can be used to create composite resources. The final product depends on the chosen data sources and the criteria used to merge them. The choice of different sources and the application of different redundancy criteria have led to the emergence of a number of different composite databases, each of which has its own particular format. The two composite protein databases of interest are NRDB and OWL.

The Non-Redundant Database

NRDB (Non – Redundant Database) (Jenuth, 2000) is built locally each night at the National Center for Biotechnology Information (NCBI). The database is a composite

of GenPept (derived from automatic GenBank CDS translations), PDB sequences, SWISS-PROT, spupdate (the weekly updates of SWISS-PROT), PIR and GenPeptupdate (the daily update of GenPept). This database is thus comprehensive and contains up-to-date protein sequence information. However, strictly speaking, NRDB is not a non-redundant database - only non-identical. This is because NRDB uses a rather simplistic approach to merging the primary databases. This leads to a number of problems: multiple copies of the same protein are retained in the database as a result of polymorphisms and/or minor sequence errors; incorrect sequences that have been amended in SWISS-PROT are reintroduced when retranslated from raw DNA sequences, etc. As a result, the contents of NRDB are both error-prone and, in spite of the name, redundant.

The OWL Database

OWL is a non-redundant protein sequence database built at the University of Leeds in collaboration with the DaresBury Laboratory in Warrington (Bleasby et al., 1994). The database is a composite of four major primary sources: SWISS-PROT, PIR1-4, GenBank (CDS translation) and NRL-3D. During the amalgamation procedure, both identical copies of sequences and those containing single amino acid differences are eliminated, thus leading to a compact and efficient resource for sequence comparisons. The OWL database can be downloaded from the NCBI ftp server.

Sequence Comparison

Over the past 20 years, sequence comparison has evolved from an obscure pursuit of few evolutionary biologists to a routine event that is performed 100,000's times a day

by more than 10,000 different scientists in 100 different countries. This is because sequence comparison is the simplest, quickest and most inexpensive way of determining whether a new gene or protein might do something interesting. By comparing a sequence to others that have already been painstakingly characterized, it is possible to suggest not only functional and structural similarity, but also detailed phylogenetic relationships – simply on the basis of sequence similarity alone.

Algorithms for Sequence Comparison

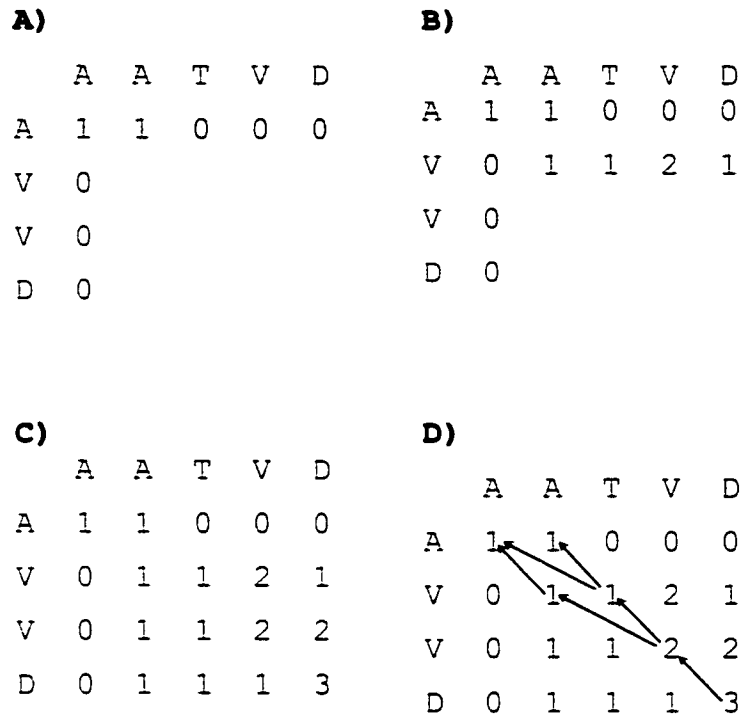
One of the most common methods for performing sequence alignment involves a technique called dynamic programming (Sankoff, 1983; Waterman, 1984; Pearson and Lipman, 1988). Dynamic programming is an efficient mathematical technique that can be used to find optimal “paths” or routes to multiple destinations. Dynamic programming is also useful in locating paths that could be combined to achieve some maximum score. The application of dynamic programming to sequence alignment was first illustrated by Needleman and Wunsch in 1970. In this now classic paper (Needleman and Wunsch, 1970), these authors demonstrated how dynamic programming actually permits a quantitative assessment of sequence similarity, while at the same time showing how two sequences can be globally aligned (Figure 1.2). The recursive function that is used in scoring is written as follows:

$$S_{i,j} = S_{i,j} + \max \left\{ \begin{array}{l} S_{i-1,j-1} \\ \text{Max } S_{i-x,j-1} + W_x \text{ (} 2 < x < i \text{)} \\ \text{Max } S_{i-1,j-y} + W_y \text{ (} y < 2 < i \text{)} \end{array} \right\} \quad (1.1)$$

$S_{i,j}$ is the score for the alignment ending at i in sequence 1 and j in sequence 2

W_x is the score for making a x long gap in sequence 1

W_y is the score for making a y long gap in sequence 2



The best alignments:

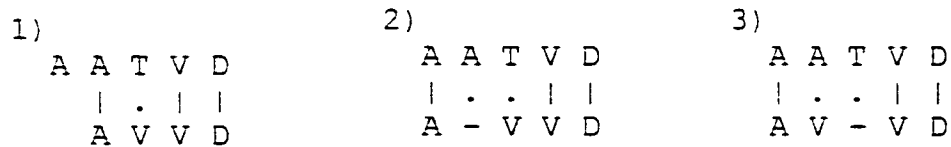


Figure 1.2 Illustration of how dynamic programming can be used to calculate a simple sequence alignment. In this calculation identical matches receive a score of 1, mismatches receive a score of 0 and no penalty is applied for gaps. The recursive function shown in Equation 1.1 is used to fill out the alignment matrix.

A. After calculating the first row and the first column of the score matrix.

B. After calculating the second row and the first column of the score matrix.

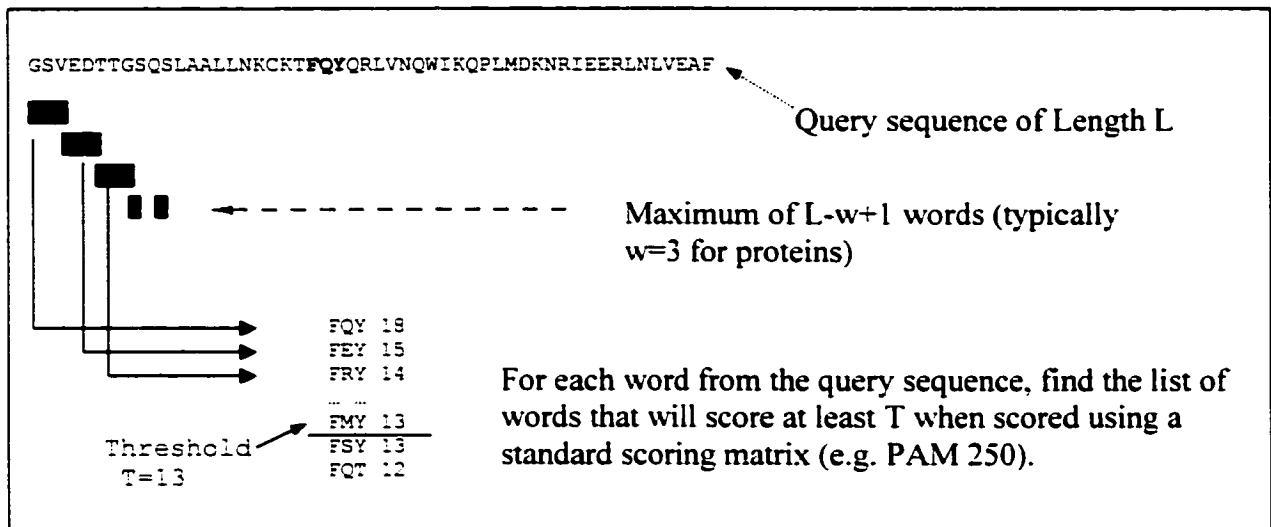
C. After calculation of the last position S4,5 in the score matrix and completed path graph.

D. The path graph is shown as arrows indicating the best alignments.

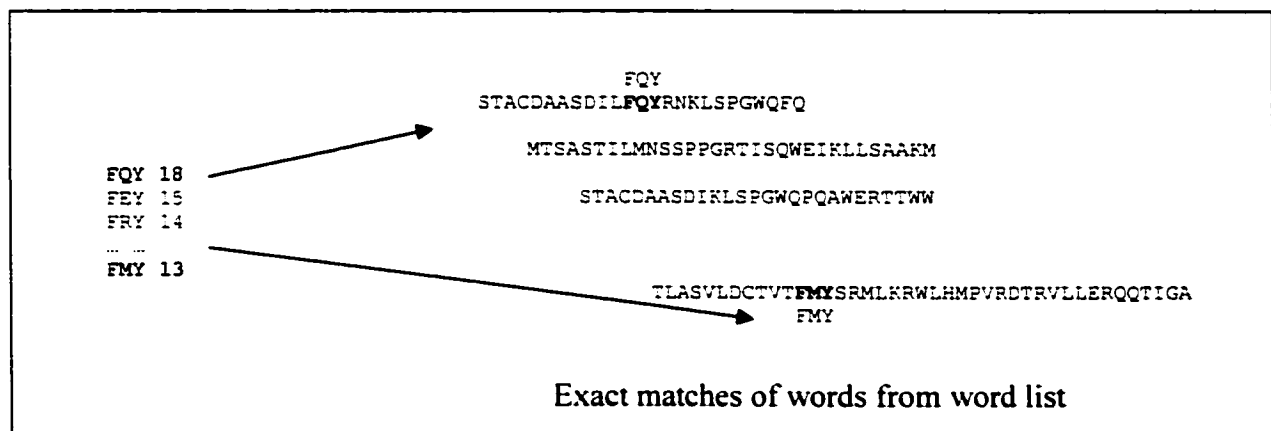
Dynamic programming is mathematically the most rigorous method for determining a pairwise sequence alignment. However, it is inherently slow. Dynamic programming is known in the computer world as an $O(N^2)$ algorithm, meaning that every time you double the length of the two sequences you are comparing, the time to analyze them will increase by a factor of four. If one wished to perform 800,000+ comparisons using this approach (the size of a typical database search these days), it could easily take several hours on a fast (1.4GHz) computer. Given that most people do not want to tie up a computer for that long or wait hours for an answer, there has been a great deal of effort directed at developing methods to improve search speeds so that database comparisons could be done more quickly. However, improvements in speed usually end up sacrificing accuracy or precision. Nevertheless, the advent of such fast “N-type”(calculation is linear with N, the size of the database) algorithms as FASTA and BLAST has revolutionized the process and frequency of sequence comparison and database searching.

The FASTA algorithm, first described by Lipman and Pearson (1985), is based around the idea of identifying short words, or k-tuples, common to both sequences under comparison. K-tuple sizes of 1 or 2 residues are typically used in protein searches, while larger k-tuples (up to 6 bases) are used in DNA searches. Comparison of k-tuples, and their relative offsets between the two sequences, can be viewed as focusing on diagonal matches in a dynamic programming matrix. FASTA uses a heuristic approach to join k-tuples that lie close together on the same diagonal. The regions formed in this way contain mismatches lying between matching k-tuples. If a significant number of matches are found, FASTA uses a dynamic programming algorithm to compute gapped

1) For the query sequence of length L, find the list of high scoring words of length w.



2) Compare the word list to the database and identify exact matches.



3) For each word match, extend alignment in both directions to find alignments that score greater than the threshold score S.

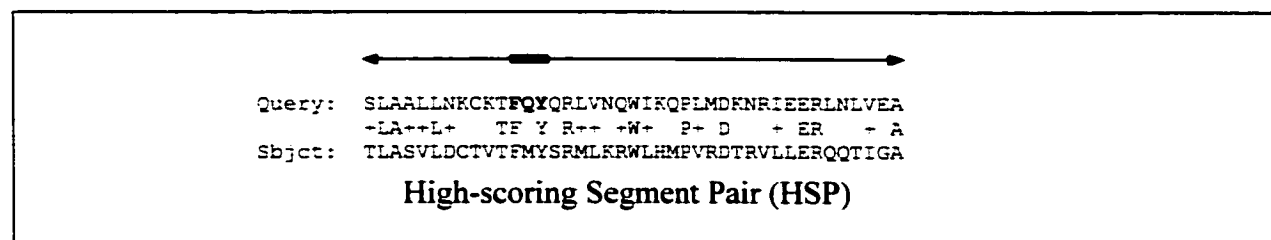


Figure 1.3 The BLAST searching algorithm

alignments that incorporate the ungapped regions.

In 1990, a second ‘fast’ algorithm for database comparison was introduced - called BLAST. BLAST is short for Basic Local Alignment Search Tool. The BLAST algorithm (illustrated in Figure 1.3) also relies on the identification of short subsequences (k-tuples), which serve as the core of an alignment. Multiple k-tuples can be combined and extended to serve as “seeds” for a more extended alignment, allowing for some number of insertions, deletions, or changes between the two sequences. BLAST pre-calculates what are called High Scoring Pairs (HSP) clusters at k-tuples deemed to be statistically significant to initiate an alignment between two sequences. Once it finds these HSPs, it looks for matching words of any length that score above a pre-set threshold. The use of statistics, larger word sizes and more sophisticated programming techniques has made BLAST even faster than FASTA. Versions of both FASTA and BLAST programs exist for comparing either a nucleic acid or protein query sequence to a database of one or the other kind of sequence.

Relative to dynamic programming methods, FASTA was shown to accelerate database searching process by a factor of 10 or more. The BLAST algorithm further accelerated database searching by a factor of 2-3. In Chapter 2, we use a hybrid FASTA/BLAST approach called FAST_ALIGN, described by Wishart et al. (1994a), to perform protein sequence alignment as part of a generalized eukaryotic for the gene prediction method.

1.5 Sequence Analysis - Exon Prediction

Every living cell contains one or more large DNA molecules called chromosomes. Each chromosome contains many genes - the basic physical and functional units of heredity. In eukaryotes, each gene is composed of at least one exon and/no introns. An exon is a specific region of DNA sequence whose sequences are translated into proteins, which provide the structural components of cells and tissues as well as the enzymes for essential biochemical reactions. The regions between exons are called introns, which have no coding function. A picture of eukaryotic gene structure is shown in Figure 1.4. Identifying coding regions is essential to identifying and understanding the functions of their encoded proteins. Given the incredibly fast rate of DNA sequence generation that is occurring today, most researchers agree that faster, better and, more accurate ways of annotating genomic sequence data must be developed. As a consequence Computational gene finding is becoming a very active area of bioinformatics research.

Gene finding systems come in two major types: eukaryotic and prokaryotic. Prokaryotic gene finding is relatively easy, since prokaryotic genomes are very small, gene rich (85-90% coding sequence) and they typically have no introns. One can be fairly successful in finding genes simply by identifying long open reading frames (ORFs). There are currently two main systems for prokaryotic gene finding, both of which are more than 95% accurate: GeneMark (Borodovsky and McIninch, 1993) and Glimmer (Salzberg, 1997).

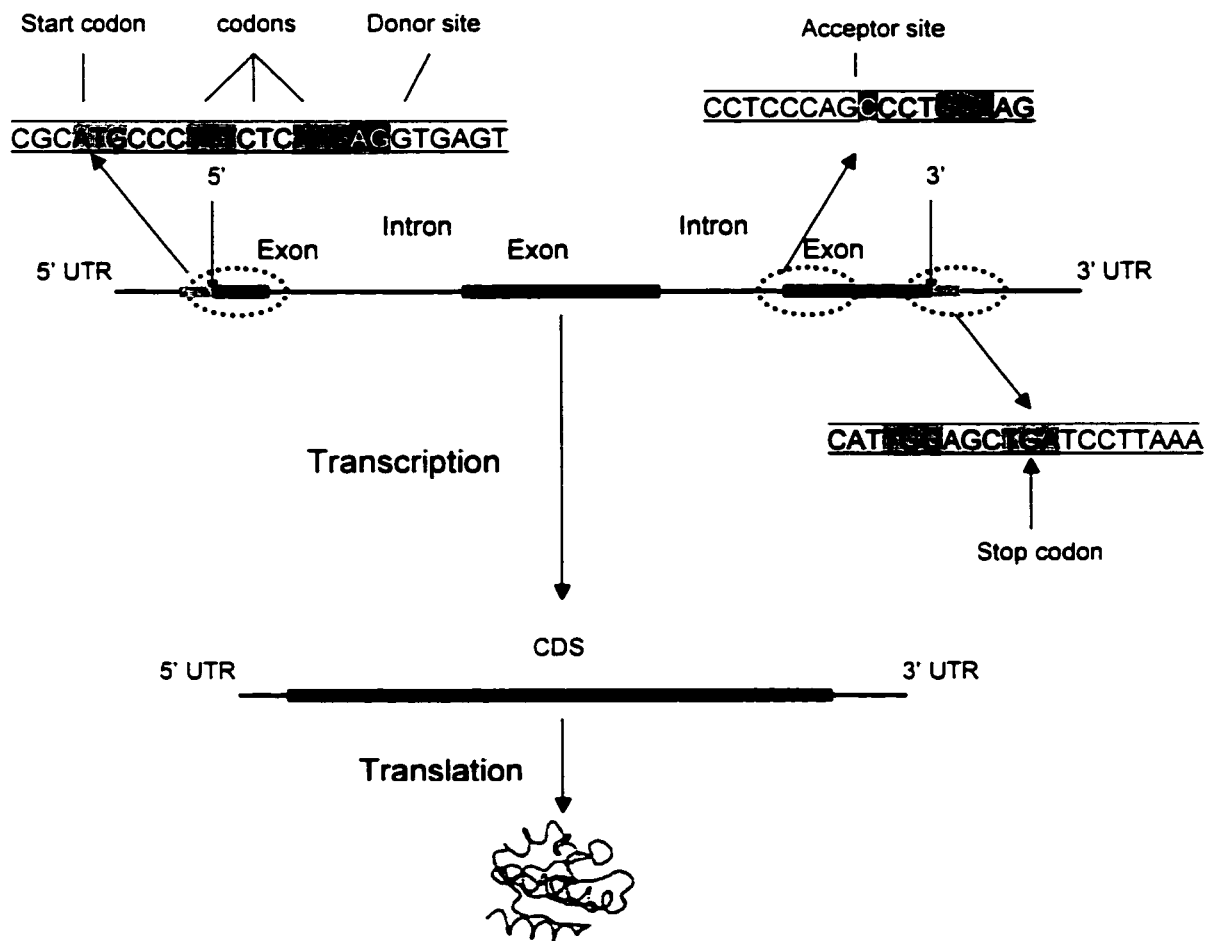


Figure 1.4 The structure of eukaryotic genes with some of the important transcriptional signals shown.

Eukaryotic gene finding is usually harder, since eukaryotic genomes are larger and more complex than prokaryotic genomes. Large-scale sequencing projects have motivated the need for a new generation of algorithms for eukaryotic gene recognition. There have been two different approaches to eukaryotic gene prediction, one is statistically based and the other is similarity based.

Statistically based methods rely on identifying signals or sequence patterns or probability profiles. The simplest statistical approach involves identifying several signals including: a) the start codon; b) a donor site (GT - the beginning of each intron); c) an acceptor site (AG - the end of each intron); and d) the stop codon. However, these simple signals and patterns are not usually enough. Codon usage, amino acid usage, periodicities in codon usage and other statistical parameters must also be used to find genes (Gelfand, 1995). For example, codon usage differs between coding and non-coding regions, thus enabling one to use this measure to frequently identify genes (Fickett, 1982; Staden and McLachlan, 1982). Gribskov et al. (1984) use a likelihood ratio approach to compute conditional probabilities, while others use hidden Markov models, decision trees, or neural networks. For example, GRAIL is a neural-net based system; HMMGene (Krogh, 1997) and GENSCAN (Burge and Karlin, 1997) are based on different types of hidden Markov models, while MORGAN (Salzberg et al., 1998) is a system based on decision trees.

Similarity-based approaches to gene prediction rely on the simple fact that a newly sequenced gene has a good chance of having an already known relative in the

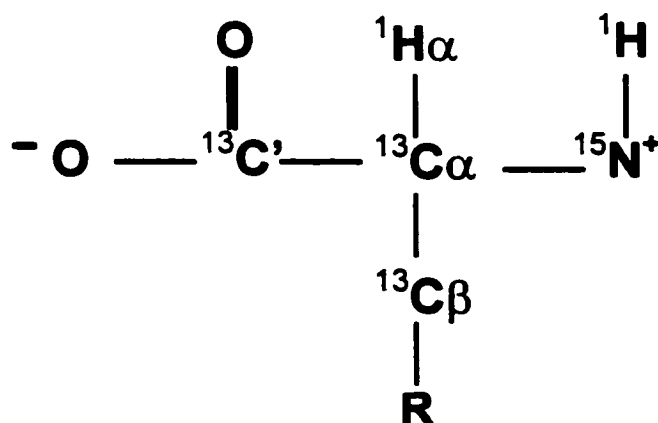
database (Bork and Gibson, 1996). The idea is based on the fact that cDNA, EST (expressed sequence tags) or protein sequences are all composed of “pure” exons. Therefore, given a genomic sequence, one can usually find a set of candidate sequence blocks that contain all true exons by performing a sequence similarity search of the cDNA, EST or protein databases. Given the fact that these sequence databases are growing exponentially, it suggests that the odds of finding a matching candidate will grow exponentially too. As a result, the trend in gene prediction in the late 1990s shifted from statistically-based approaches towards similarity-based and EST-based algorithms.

Most eukaryotic gene prediction programs today can correctly identify genes with an accuracy of 65~80% at the nucleotide level. While much better than early methods described in the late 1980's, this performance is still unsatisfactory. Chapter 2 describes a new gene identification program - GRPL+, which correctly predicts genes with an accuracy of 97%. This new approach uses a statistical method – reference point logistic classification developed by Peter Hooper (1999), in combination with sequence similarity searches to greatly improve the overall performance.

1.6 Bioinformatics Applications in NMR

Nuclear magnetic resonance (NMR) is the phenomenon that occurs when the nuclei of certain atoms are immersed in a static magnetic field and exposed to a second oscillating magnetic field. Some nuclei experience this phenomenon, and others do not, dependent upon whether they possess a property called spin. Spin is a fundamental property of nature like electrical charge or mass. The nucleus with non - zero spin are

observable by NMR. Although almost every element in the periodic table has an isotope with a non - zero nuclear spin, NMR can only be performed on isotopes whose natural abundance is high enough to be detected and who have an odd number of protons and/or neutrons. For instance, ^{12}C (the most common isotope of carbon) produces no NMR signal while ^{13}C (which is 1% abundant) produces an easily detectable NMR signal. The nuclei routinely used in NMR are ^1H , ^{13}C , ^{15}N , ^{19}F and ^{31}P . Nuclear magnetic resonance spectroscopy makes use of the NMR phenomenon to study physical, chemical, and biological properties of matter. As a consequence, NMR spectroscopy has found many applications in many areas of science. In this dissertation, I will specifically discuss NMR applications in protein chemistry. The isotopes or atoms in amino acids that are of greatest interest in protein chemistry are $^1\text{H}_\alpha$, $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$, ^{13}CO , ^{15}N , ^1HN , as shown below (R represents the amino acid side chain; $^{13}\text{CO} = ^{13}\text{C}^{\cdot}$)



The true usefulness of NMR as a technique for structural biology depends on the fact that NMR resonances are not wholly dependent on nuclear properties. Different atomic or molecular structures can lead to different absorption frequencies. These

chemically and structurally dependent changes in absorption frequency are known as chemical shifts. It has been long recognized that chemical shifts contain important structural information about for proteins. For example, proteins are known to have two characteristic subsets of secondary structure – helices and beta strands. The $^1\text{H}\alpha$ chemical shift for most residues is different enough between helices and beta strands to manifest itself as either an upfield shift (for helices) or a downfield shift (for beta-strands) (Wishart and Sykes, 1994b). These chemical shift changes have also be observed for ^{13}C , ^{15}N nuclei of amino acids as well (Le and Oldfield, 1994; Wishart and Sykes, 1994c). By plotting the difference between the chemical shifts arising from the secondary structure and the chemical shifts expected in a random coil (the so-called secondary chemical shift), one can get a simple plot which shows the location and length of helices and beta-strands in a protein or polypeptide – called the chemical shift index (Wishart et al., 1992).

Technical improvements in high resolution NMR spectroscopy since the 1980s, combined with enhanced molecular biology techniques, have provided powerful tools to study the structure and dynamics of macromolecules (Wuthrich, 1986; Markley, 1990; Stockman and Markley, 1990). Two and three-dimensional NMR techniques are now frequently used, and are playing an increasingly important role in the structure determination of protein, DNA or RNA molecules. These efforts have led to a proliferation of assigned macromolecular chemical shift data. However, due to space limitations and cost overruns, journal editors have long been under pressure to reject or avoid publishing chemical shift data. In response to this issue, a biomolecular chemical shift database – the BioMagResBank (BMRB) (Seavey et al., 1991) was set up in 1990 to

archive chemical shift data (<http://www.bmrb.wisc.edu>). The number of protein chemical shifts in the BMRB has grown from 20,000 at the time of its launch to more than 560,000 protein chemical shifts now. To date, NMR spectroscopy has been used to solve more than 2500 biomolecular structures, many of which are now deposited in the Protein Data Bank (Bernstein, et al., 1977) and Nucleic Acid Database (NDB) (Berman et al., 1992).

When the BMRB became available, researchers began to use it to develop related bioinformatics tools for various NMR-related tasks. SHIFTY (Wishart et al., 1997) and TALOS (Cornilescu et al., 1999) are examples of two bioinformatics applications in the NMR area. The purpose of both SHIFTY and TALOS is to simplify the process of NMR chemical shift assignment or structural refinement.

The primary goal in biomolecular NMR is to assign as many chemical shifts as possible for each residue in a protein sequence. However, most resonance assignment schemes depend on the spectroscopist having some knowledge of approximate chemical shifts or expected chemical shift ranges for the residues under question. Assigning chemical shifts for each residue in a protein sequence is usually a time-consuming task, even for an expert. Inspired by the fact that homologous sequences exhibit not only similar structures but the similar secondary chemical shifts as well (Redfield and Robertson, 1991), Wishart and colleagues (1997) developed an automated protein chemical shift prediction program – SHIFTY, to help assign chemical shifts for proteins. The technique uses dynamic programming to detect sequence homology between the query protein and the sequences of hundreds of previously assigned proteins in the

BMRB. Once a homolog is found, SHIFTY uses a simple set of rules to directly assign or transfer a complete set of ^1H , ^{13}C or ^{15}N chemical shifts to the unassigned protein. In this way, SHIFTY makes the sequential assignment process substantially easier, significantly faster and much less dependent on NOEs or other information. The SHIFTY web server (<http://redpoll.pharmacy.ualberta.ca/shifty>) has been linked to BMRB web server and widely used by NMR spectroscopists.

TALOS is another example of how bioinformatics techniques can be applied to NMR spectroscopy. TALOS (Cornilescu et al., 1999) uses sequence similarity to effect backbone dihedral prediction, with an accuracy of $\pm 15^\circ$. Specifically, TALOS compares a query protein (and its associated chemical shifts) to a database of previously assigned proteins, including their sequence, their chemical shifts and their corresponding backbone dihedral angles (as determined by X-ray crystallography). TALOS uses a very simple measure of sequence similarity to predict the most likely backbone dihedral angles from homologous peptides (based on a combined measure of sequence similarity and chemical shift similarity). In this way TALOS offers a simple, intuitive approach to converting raw chemical shift information into useful structural restraints for NMR-based structure generation and refinement.

As discussed above, more and more biomolecular NMR spectroscopists are interested in interpreting chemical shifts in peptide and protein NMR studies. However, past problems and inconsistencies in NMR data collection methods have led to a growing problem but previously undetected for chemical shift assignments. Indeed, about 20~30%

chemical shifts deposited in the BMRB are either mis-assigned or mis-referenced (see Chapter 3). In Chapter 3, we describe the development of a corrected protein chemical shift database – RefDB. RefDB is a self-updating, self-correcting database of carefully corrected or re-referenced chemical shifts, derived from the BMRB. The process involves predicting protein ^1H , ^{13}C and ^{15}N chemical shifts using X-ray or NMR coordinate data via a specially developed program called SHIFTX (Neal and Wishart, manuscript in preparation), and then comparing those predictions to the observed shifts reported in the BMRB (via SHIFTCOR). RefDB provides a standard chemical shift resource for NMR spectroscopists, wishing to derive or compute chemical shift trends in peptides and proteins.

1.7 My Contribution to This Dissertation

This dissertation contains contributions from a number of people, as indicated in the acknowledgment page. In particular, I would like to thank Dr. Peter Hooper for the development and implementation of GRPL in chapter 2. I would also like to thank Steve Neal and Dr. Alexander Nip for their contributions to SHIFTX which is described in chapter 3.

Concerning my contributions to the work described in chapter 2, I was responsible for almost all aspects of the data collection, reduction and analysis. This included collecting the test and training data sets from GENBANK, performing statistical analyses for the different gene structure prediction programs, developing a program to

calculate all of the quoted statistical values and finally developing, implementing and testing the protein sequence alignment algorithm and program for GRPL+.

In chapter 3, I developed two major programs – SHIFTCOR and UPDATE. These two programs automated the process of auto-correcting and auto-updating the RefDB database. This process includes the data retrieval, data evaluation and statistical analysis between the observed and predicted chemical shift data. I was also responsible for setting up the RefDB web server, installing the GLIMPSE search engine, the BLAST sequence searching tool and implementing all of the related CGI programs. I was also responsible for performing the manual checking and data verification of the RefDB database.

Chapter 2*

Prediction of Genetic Structure in Eukaryotic DNA using Reference Point Logistic Regression and Sequence Alignment

2.1 Introduction

The Human Genome Project, along with various projects in the pharmaceutical, agricultural, and forestry industries, is creating an enormous quantity of raw, unannotated DNA sequence data. With this abundance of data, there is a growing need for more effective software tools to extract vital information from this raw data. Tools for identifying protein coding regions and predicting complete genes are of particular importance. Since the early 1990's, a number of computer programs for eukaryotic gene identification have been developed, tested, and described in the literature. These include: SORFIND (Hutchinson and Hayden, 1992), GeneID (Guigo et al., 1992), GENMARK (Borodovsky et al., 1993), Xpound (Thomas and Skolnick, 1994), FGENEH (Solovyev et al., 1994; Solovyev and Salamov, 1997), GRAIL2 (Xu, 1994), GeneParser (Snyder and Stormo, 1995), Genie (Kulp, 1996), and GENSCAN (Burge and Karlin, 1997). Most of these programs make use of sophisticated pattern recognition techniques such as linear discriminant analysis, neural networks, or Hidden Markov models to identify coding regions. Some programs also make use of database sequence alignment methods, such as BLAST or XBLAST, to further improve their predictions (Kneche, 1995; Searls, 1995; Snyder and Stormo, 1995). The idea of using a similarity-based approach to gene

*Portions of this chapter were published as a paper by Hooper, P.M., Zhang, H. & Wishart, D.S. in *Bioinformatics* 16(5), 425-438 (2000).

detection was first proposed by Gish and States (1993). The expectation was that previously identified homologs (genes or proteins) could be used to better define exon/introns boundaries and to correct possible reading frame errors. Therefore, information from homologous sequences can be used not only for gene detection, but also for detailed prediction of exon-intron structure as well.

The evaluation of gene prediction accuracy has always been an area open to much controversy and debate (Burset and Guigo, 1996). Different evaluation methods will reveal different strengths and weaknesses for different methods. The difficulty in evaluating gene prediction arises from the fact that it is not simply a two state (right/wrong) problem. There are actually four states to consider: true positives (TP), false positives (FP), true negative (TN) and false negatives (FN). Depending on how one wants to bias their prediction one can always develop a method to preferentially eliminate false positives (improve specificity) or reduce the number of false negatives (improve sensitivity). Ideally one would want to improve both sensitivity (S_n) and Specificity (S_p) simultaneously. Indeed, a perfect prediction always has a sensitivity and specificity of 1.0. How then is gene prediction accuracy normally evaluated? In Figure 2.1 we provide a simple illustration of the meaning of at least some these definitions and scoring schemes. Hopefully this should help in later discussions in this chapter. However, it is important to note that this figure illustrates the common evaluation methods for gene prediction at the single base or single nucleotide level only. Because introns, exons and intergenic regions typically consist of hundreds or thousands of consecutive bases, information about the start, end and length of these regions should ideally be encoded into any evaluation

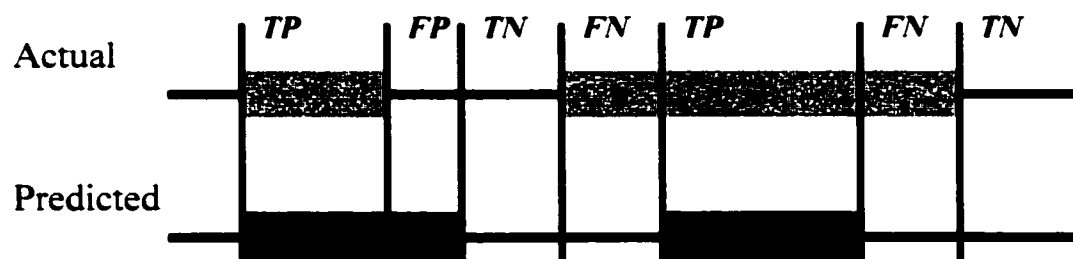


Figure 2.1 Evaluation statistics for gene prediction. TP = True Positive. FN = False Negative. TN = True Negative. FP = False Positive.

Sensitivity	Fraction of actual coding regions that are correctly predicted as coding $S_n = TP / (TP + FN)$
Specificity	Fraction of the prediction that is actually correct $S_p = TP / (TP + FP)$
Correlation	Combined measure of sensitivity and specificity $CC = (TP * TN - FP * FN) / [(TP + FP)(TN + FN)(TP + FN)(TN + FP)]^{0.5}$

procedure. This has led some groups to prefer to evaluate prediction accuracy not at the base or nucleotide level, but at the exon/intron level. When this more rigorous approach is adopted, the accuracy of many gene prediction methods falls to embarrassingly low levels.

Because of the continuing controversies over evaluation procedures Burset and Guigo (1996) compared many of these programs with a large test set of 570 vertebrate sequences using several accuracy measures. Among those methods not using database searches, the average correlation coefficient (CC) at the nucleotide level varied from 0.65 to 0.80 while the combined average specificity and sensitivity (Avg) at the exon level varied from 0.17 to 0.63. Among those methods that used database searches as an adjunct to exon prediction, their CC values ranged from 0.85 to 0.87 while their Avg ranged from 0.57 to 0.71. To date the best results reported for exon prediction belong to the GENSCAN program described by Burge and Karlin (1997). This method, which does not use explicit database alignments, yields a CC of 0.92 and an Avg of 0.80 when tested on the dataset of Burset and Guigo. These results suggest that, while definite progress has been made, there is still some room for improvement.

In this chapter we describe a novel approach to gene prediction in eukaryotic organisms. This technique makes use of a new to statistical classification technique called referencing point logistic (RPL) developed by Dr. Peter Hooper. Reference point logistic (RPL) regression (Hooper, 2001) is a generalization of logistic regression (Cox, 1989) that can be used in complex classification problems to model the conditional probability

that an item belongs to a specified class given features observed for the item. RPL regression is closely related to a classification technique in which reference points are used to construct piecewise linear classification boundaries (Hooper, 1999). An approximated idea of how the method works can be gained by looking at Figure 2.2. This figure shows some artificial data for a two-group classification problem with two features. The piecewise linear boundary between the two groups represents an RPL classifier based on four reference points, two for each class. The linear boundary segments are perpendicular to lines joining pairs of reference points (not shown in the plot). The position of the linear segments, relative to the reference points, is controlled by additional parameters. The reference points and additional parameters are determined using a training algorithm similar to a neural network backpropagation algorithm (Michie, 1994). A corresponding RPL regression model represents the conditional probability of a class given the observed features as a smooth function of the location in the plot. Therefore, the estimated probability for a point in class A is close to 1.0 at points well below the boundary, close to 0.0 at points well above the boundary, and close to 0.5 at points on the boundary.

Our gene prediction program has two stages. In the first stage, RPL regression models are used to calculate scores for potential functional sites at exon boundaries. These are combined with scores for interval content, length, and state (via a Generalized Hidden Markov Model) to determine a score for each possible parse of a sequence into exons, introns, and intergenic regions. An optimal parse is then found using a dynamic programming algorithm (Needleman and Wunsch, 1970). In the second stage, protein

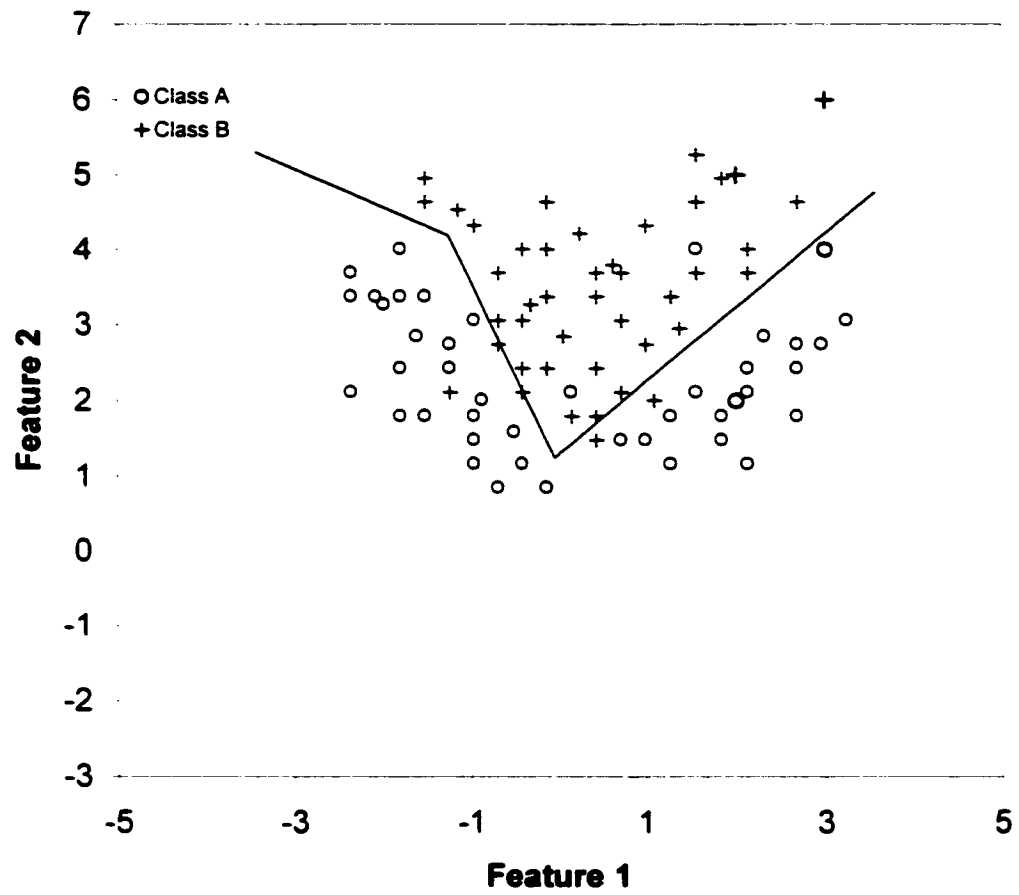


Figure 2.2 Illustration of RPL classification boundaries. Plot symbols denote the actual classes. The RPL classifier would assign points below the boundary to class A.

sequence alignment versus OWL protein database (version 29.4) methods are applied to improve the accuracy of the parse. Overall we found the results of this second step to be quite promising and show that protein sequence alignment can be and should be a routinely used to improve upon initial exon prediction results. We refer to the first stage of the program as GRPL (pronounced grapple) and to the full two stage program as GRPL+. In this chapter, we mainly describe the methods and results associated with GRPL+ and how it improves GRPL predictions. A more complete description of GRPL can be found in Hooper et al. (2000). GRPL and GRPL+ are both capable of predicting the genetic structure of vertebrate, invertebrate, and plant DNA with an accuracy exceeding that of other programs that were tested on the same data set. We also report on how the size of the protein database and the accuracy of the initial GRPL predictions affect the accuracy of the final GRPL+ predictions.

2.2 Methods

Training and Test Sets

Three versions of GRPL and GRPL+ were evaluated, trained on separate sets of human, *Drosophila*, and *Arabidopsis* sequences. The human version of GRPL (GRPL (Hu)) was trained on 367 human DNA sequences. These were selected from 380 sequences (238 multi-exon genes and 142 single-exon genes) compiled by Burge & Karlin (1997). We dropped 13 of the single-exon genes where the translation initiation and/or termination sites were too close to the end of the sequence to be useful for modeling the functional sites. When estimating the coding region model for GRPL (Hu), the training set was augmented by 1618 human cDNA sequences. These were selected

from 1619 human cDNA sequences compiled by Burge & Karlin (1997). One cDNA sequence (HSCA2VR) was dropped because it did not begin with ATG or end with a stop codon.

We assembled a set of 171 *Drosophila* DNA sequences from the Oct. 1998 version of GenBank (Benson et al., 2000). The *Drosophila* version of GRPL (GRPL (Dr)) was trained on the first 139 sequences, sorted by name, and the remaining 32 sequences were assigned to a test set. Fourteen sequences (10 in the training set and 4 in the test set) contain nonconsensus splice sites (21 of the 5' splice sites and 15 of the 3' splice sites). We suspect that most of the nonconsensus sites are due to data entry or annotation errors. There is usually a consensus site in close proximity to the reported site. The nonconsensus sites appear to have little effect on prediction accuracy measured at the nucleotide level. The effect at the exon level is more substantial, since GRPL does not predict exons with nonconsensus sites. We also assembled a set of 272 *Arabidopsis* DNA sequences from GenBank (October, 1998), where all splice sites were consensus sites. A test set was created with the first 32 sequences, sorted by name, and GRPL (Ar - *Arabidopsis*) was trained on the remaining 240.

Three additional test sets were used: the 570 vertebrate DNA sequences assembled by Burset & Guigo (1996), and the two test sets (28 and 34 human DNA sequences) originally assembled to evaluate performance of GeneParser by Snyder & Stormo (1995). We obtained more recent versions of the GeneParser test sequences from

GenBank (October, 1998). There appear to be minor changes in several of these sequences.

Database Sequence Alignment in GRPL+

It has previously been shown that comparisons of predicted exons with protein sequence database can improve both the sensitivity and specificity of the overall prediction (Guigo et al., 1992; Gelfand, 1995; Snyder and Stormo, 1995). Based on these earlier results, we implemented a database search component into GRPL to serve as a final "knowledge-based" refinement stage. Predicted exon locations, including the predicted exon start, exon end, and reading frame, were first obtained from GRPL. The predicted exons were then spliced and translated to create a single tentative protein sequence. This initial sequence was searched against the OWL protein database (Bleasby et al., 1994) using a slightly modified form of the FAST-ALIGN program (Wishart et al., 1994a). OWL (release 29.4) is a non-redundant protein sequence database containing 198,742 peptide and protein sequences. FAST-ALIGN is a global alignment algorithm that uses n-tuple comparisons, similar to FASTA (Pearson and Lipman, 1988) and BLAST, to identify initial sequence matches followed by a global alignment using dynamic programming (Needleman and Wunsch, 1970) with generous gap insertion and extension penalties.

After the initial database alignment/scoring was complete, those sequences with global alignment scores exceeding an empirically determined cutoff value (Abagyan and Batalov, 1997) were kept for further analysis. To simulate the situation where the query sequence is not yet deposited in the database, we also removed the top scoring sequence

from each list, as this sequence often matched the query sequence identically. If no sequence homologue was found in the first pass, a second search was performed wherein each exon was translated (all three reading frames) and searched against the OWL database using the same scoring and selection criteria. If this follow-up search also failed to identify a significant match in the database, the original GRPL prediction was kept without further modification.

Once a “second-best” protein homologue was identified from the translated exons, a second pairwise comparison was performed. Specially, the protein homologue would be aligned against all three translated reading frames of the GRPL predicted (spliced) gene using a standard Needleman-Wunsch alignment algorithm. The significance of each alignment was assessed using the same empirically derived cutoffs as before. Predicted exons were then extended, shortened, linked, or combined to more closely match the database protein sequence. This three-frame comparison also allowed rapid identification and correction of mistaken GRPL predictions, indels (internal deletions), or frameshift errors. If a portion of protein sequence was found to be missing, the region of the gene that mapped to the missing segment would be translated in all three frames and a pairwise alignment performed with the corresponding segment from the database sequence to identify the missing exon or exon fragment. To remove any alignment bias that might be introduced by the presence of remotely related sequences, the final refinement step was performed only if the global pairwise sequence identity between the database homologue and the translated query sequence exceeded 40%. The time required to complete the database search and all requisite translation, alignment and comparison steps in GRPL+

is about 250 CPU seconds (180 MHz SUN SPARC5) for a typical query sequence of 300 residues.

2.3 Results and Discussion

Accuracy Measures

The accuracy of gene structure predictions was evaluated on test sets as follows. For each sequence in the test set, the predicted exons were compared with the annotated exons (GenBank "CDS" key). Standard measures of predictive accuracy per nucleotide and per exon were calculated for each sequence and averaged over all sequences for which they were defined. For a given sequence, TP (true positive) is defined as the number of nucleotides correctly predicted to be in coding regions, TN (true negative) is defined as the number of nucleotides correctly predicted to be in non-coding regions, FP (false positive) is defined as the number of nucleotides incorrectly predicted to be in coding regions, and FN (false negative) is defined as the number of nucleotides incorrectly predicted to be in non-coding regions. These four values can be arranged in a 2 x 2 array. The row and column totals are: PP = TP + FP (predicted positive), PN = TN + FN (predicted negative), AP = TP + FN (actual positive), and AN = TN + FP (actual negative). There are four ratios of potential interest. Sensitivity is defined as $S_n = TP/AP$. Specificity is usually defined as TN/AN , but the alternative definition TP/PP has become standard in the gene structure prediction literature (Burset and Guigo, 1996). We argue that the former definition has merit when considering the effect of coding proportion, but we adopt the gene prediction definition $S_p = TP/PP$ to avoid confusion. We define $S_q = TN/AN$ and $S_r = TN/PN$. Note that (S_q, S_r) is equivalent to (S_n, S_p) with the roles of

coding and non-coding regions interchanged. Two additional measures are reported: the correlation coefficient

$$CC = ((TP)(TN) - (FP)(FN)) / ((PP)(PN)(AP)(AN))^{1/2} \quad (2.1)$$

and the approximate correlation $AC = -1 + 2(Sn + Sp + Sq + Sr)/4$. Usually CC and AC have similar values but, if one of the four ratios has zero denominator, then CC is undefined while AC is redefined using the average of the other three ratios (Burset and Guigo, 1996).

Two additional measures were calculated at the exon level. XTP (exon true positive) is defined as the number of actual exons that exactly match predicted exons. Exon sensitivity Xsn is XTP divided by the number of actual exons. Exon specificity Xsp is XTP divided by the number of predicted exons.

Comparisons with Other Programs

There is growing interest in the analysis of invertebrate and plant DNA. Since genetic structure varies substantially among organisms, any gene prediction program can be improved by tailoring its parameters for specific families of organisms. GENSCAN (Burge and Karlin, 1997), Genie (Kulp, 1996), and GeneID (Guigo et al., 1992) provide options for organism type. Dr. Hooper developed three versions of GRPL, with human, *Drosophila*, and *Arabidopsis* training sets, and evaluated them using five test sets. Detailed descriptions of the training and test sets are provided in the Methods section.

When carrying out sequence alignment in GRPL+, the second best homologue match was used in order to simulate a situation where each gene sequence is novel and not yet deposited in GenBank. Usually the best homologue match was the translated sequence for the gene being analyzed.

Table 2.1 Performance comparisons for GeneParser Test Set I (28 human genes, average coding proportion 0.14).

Program	Sn	Sp	Sq	CC	AC	Xsn	Xsp
GRPL	0.96	0.88	0.978	0.90	0.91	0.72	0.69
GRPL+	0.98	0.95	0.994	0.96	0.96	0.73	0.80
GenScan	0.97	0.88	0.982	0.91	0.91	0.74	0.73
Genie	0.86	0.81	0.964	0.80	0.80	0.68	0.64
Grail 2	0.91	0.86	0.980	0.86	0.87	0.48	0.44
GeneID	0.80	0.84	0.979	0.79	0.79	0.60	0.51
Xpound	0.76	0.87	0.984	0.78	0.79	0.21	0.24

Table 2.1 contains performance results using the Burset/Guigo set of vertebrate sequences. Results, other than those for GRPL, were obtained from published sources. GRPL (Hu) denotes the version of GRPL trained on human DNA. Results for GRPL (Hu) match those of GENSCAN at the nucleotide level, and are slightly worse at the exon level. Results for GRPL (Hu)+ are substantially better at both levels.

Table 2.2 Performance comparisons for GeneParser Test Set II (34 human genes, average coding proportion 0.17).

Program	Sn	Sp	Sq	CC	AC	Xsn	Xsp
GRPL	0.89	0.93	0.989	0.89	0.90	0.69	0.77
GRPL+	0.93	0.96	0.995	0.94	0.94	0.71	0.76
GenScan	0.89	0.92	0.986	0.90	0.89	0.68	0.69
Genie	0.75	0.76	0.967	0.71	0.72	0.54	0.51
Grail 2	0.70	0.81	0.978	0.73	0.72	0.36	0.32
GeneID	0.70	0.75	0.961	0.68	0.71	0.49	0.48
Xpound	0.71	0.91	0.987	0.76	0.78	0.26	0.27

Tables 2.2 through 2.5 contain performance results based on smaller test sets. Results for Xpound were obtained by running the Xpound program locally. Results for GENSCAN, Genie, Grail2, and GeneID were obtained by submitting test sequences to their corresponding web servers^a. Tables 2.2 and 2.3 contain performance results using the GeneParser test sets of human DNA sequences, described in Snyder & Stormo (1995). The 'organism' option was set to vertebrate or human in GENSCAN, GeneID, and Genie, and the default in Grail2 and Xpound. Results for GRPL (Hu) and GENSCAN are again similar. Our findings differ in some respects from those of Burge & Karlin (1997, see Table 2.2). The differences may be related to several factors: possible improvements in some of the programs being tested, minor changes in some of the GenBank sequences, and a different convention used for averaging statistics.

Table 2.3 Performance comparisons for *Drosophila* test set (32 genes, average coding proportion 0.45).

Program	Sn	Sp	Sq	CC	AC	Xsn	Xsp
GRPL	0.98	0.96	0.975	0.94	0.94	0.68	0.68
GRPL+	0.98	0.99	0.992	0.97	0.97	0.70	0.77
GenScan	0.98	0.95	0.951	0.93	0.93	0.70	0.70
Genie	0.90	0.95	0.982	0.88	0.88	0.59	0.64
Grail 2	0.83	0.91	0.942	0.75	0.76	0.17	0.16
GeneID	0.79	0.97	0.986	0.82	0.76	0.47	0.57
Xpound	0.86	0.89	0.903	0.75	0.76	0.11	0.18

^aWeb sites:

GENSCAN server: <http://gnomic.sta.nford.edu/GENSCANW.html>

Genie server: <http://www-hge.ibi.gov/projects/genie.html>

Grail server: <http://compbio.orni.gov/Grail-bin/EmptyGrailForm>

GeneID-3 server: <http://apolo.imim.es/genelid.html>

Xpound program: <ftp://igs-server.cnrs-mrs.fr/pub/Banburg/xpound/>

Table 2.4 contains performance results for a test set of 32 *Drosophila* sequences. The human and *Drosophila* versions of GRPL were tested here. The ‘organism’ option was set to vertebrate in GENSCAN and GeneID, *Drosophila* in Genie, and the default in Grail2 and Xpound. Results for GRPL (Dr) and GENSCAN are similar, and only slightly better than those for GRPL (Hu). The presence of nonconsensus splice sites in four of the test sequences (see the Methods section) adversely affected performance at the exon level for all methods.

Table 2.4 Performance comparisons for Arabidopsis Test set (32 genes, average coding proportion 0.43).

Program	Sn	Sp	Sq	CC	AC	Xsn	Xsp
GRPL	0.97	0.92	0.943	0.90	0.90	0.80	0.74
GRPL+	0.98	0.93	0.958	0.92	0.92	0.80	0.80
GenScan	0.92	0.91	0.939	0.86	0.86	0.72	0.71
Genie	0.27	0.76	0.966	0.30	0.32	0.14	0.18
Grail 2	0.39	0.87	0.968	0.44	0.46	0.13	0.10
GeneID+	0.64	0.86	0.945	0.61	0.61	0.46	0.45
Xpound	0.23	0.79	0.977	0.29	0.33	0.01	0.02

Table 2.5 contains performance results for a test set of 32 *Arabidopsis* sequences. The human and *Arabidopsis* versions of GRPL were tested here. The ‘organism’ option was set to *Arabidopsis* in GENSCAN, human or other in Genie, plants in GeneID, and the default in Grail2 and Xpound. While GRPL (Hu) performed reasonably well here, GRPL (Ar) did substantially better. It appears that the use of a specialized training set is more effective for *Arabidopsis* than for *Drosophila*. The improvement in GRPL (Ar) over GRPL (Hu) in Table 2.5 is greater than the improvement in GRPL (Dr) over GRPL (Hu) in Table 2.4. This result may be related to differences in average C+G content: 49% for

the Burst/Guigo test set, 53% and 52% for the Gene Parser test sets, 47% for the *Drosophila* test set, and 39% for the *Arabidopsis* test set.

Table 2.5 Performance comparisons for the Burset/Guigo test set (570 vertebrate genes, average coding proportion 0.21). GENSCAN results are from Burge & Karlin (1997). Genie results are from Kulp et al. (1996). The remaining results are from Burset & Guigo (1996).

Program	Sn	Sp	Sq	CC	AC	Xsn	Xsp
GRPL	0.93	0.93	0.984	0.91	0.91	0.76	0.79
GRPL+	0.97	0.97	0.990	0.96	0.96	0.81	0.85
GenScan	0.93	0.93	n/a	0.91	0.92	0.78	0.81
Genie	0.76	0.77	n/a	0.72	N/A	0.55	0.48
Grail 2	0.72	0.87	n/a	0.75	0.76	0.36	0.43
GeneID+	0.91	0.91	n/a	0.88	0.88	0.73	0.70
Xpound	0.61	0.87	n/a	0.68	0.69	0.15	0.18
GeneID	0.63	0.81	n/a	0.65	0.67	0.44	0.46
GeneParser3	0.86	0.91	n/a	0.85	0.86	0.56	0.58

The comparisons reported above are for sequences containing a single complete gene. We carried out a small additional study based on four multigene sequences: Z83317 (*Caenorhabditis elegans*, 2 genes, 12 exons, 32kb), Z47352 (mouse, 4 genes, 7 exons, 14kb), Z38015 (mouse, 1 full and 1 partial gene, 17 exons, 12kb), and AB008545 (*Schizosaccharomyces pombe*, 2 genes, 4 exons, 4kb). Our results using GRPL (HU) were: Sn = (0.93, 0.95, 0.94, 1.00) and Sp = (0.70, 0.56, 0.95, 1.00). Our results using GENSCAN were: Sn = (0.65, 0.83, 0.94, 0.96) and Sp = (0.52, 0.50, 0.97, 1.00).

Effect of C+G Content

It has been shown in a number of previous studies (Lopez et al., 1994; Xu, 1994; Snyder and Stormo, 1995; Burset and Guigo, 1996) that the accuracy of gene prediction

programs tends to increase with the proportion of C+G content. Some programs, such as GENSCAN, compensate for this problem by grouping sequences into distinct C+G ‘rich’ and ‘poor’ categories. By adjusting the program parameters on the basis of these categories, it is possible to improve their overall performance while maintaining accuracy across C+G isochors (Burge and Karlin, 1997). Rather than arbitrarily grouping sequences, we have found that a continuous model of C+G content can be used to good effect. The interval content, length and state scores used by GRPL shows stability across isochors similar to that of GENSCAN. Burge and Karlin (1997) evaluated accuracy measures on four subsets of the Burset/Guigo sequences determined by proportion of C+G content: (<0.40, 0.40-0.50, 0.50-0.60, >0.60). Their GENSCAN CC averages were (0.93, 0.91, 0.92, 0.90). Our corresponding GRPL (Hu) CC averages were (0.90, 0.92, 0.90, 0.92). One might expect lower Sp levels for lower C+G content, since C+G content and coding proportion CP are positively correlated. The apparent absence of such an effect may be due to the relatively weak association between CP and C+G content within the Burset/Guigo test set.

Effectiveness of Sequence Alignment

Tables 2.1 to 2.5 show modest improvements in GRPL+ over GRPL. More substantial improvements may be difficult to attain, given the high level of accuracy achieved by GRPL alone. To investigate the improvements attainable with sequence alignment, we looked into how the accuracy of initial GRPL predictions affects the accuracy of final GRPL+ predictions. To obtain predictions with varying levels of accuracy, we employed a damaged version of GRPL in which functional site scores and

content statistics were contaminated with random errors. Table 2.6 shows accuracy results for initial predictions (IP) and final predictions (FP) after sequence alignment. Note that improvements in Sn and Xsn are relatively small compared with improvements in Sp and Xsp. Based on these results it appears that sequence alignment is most effective in removing non-coding segments from predicted genes. Sequence alignment is less effective in adding coding segments. Predicted exons can be extended but exons that are not initially predicted by GRPL are often not recovered.

Table 2.6 Effective of sequence alignment in improving initial predictions of various accuracies. Performance is evaluated on the Buset/Guigo test set. The initial predictions (IP) were obtained using damaged versions of GRPL. Sequence alignment was then applied to obtain final predictions (FP). The initial predictions vary from GRPL(Hu) to random predictions. The performance of GRPL+ for 570 vertebrate genes with different predictions of GRPL.

Sn		Sp		CC		Xsn		Xsp	
IP	FP	IP	FP	IP	FP	IP	FP	IP	FP
0.93	0.97	0.93	0.97	0.91	0.96	0.76	0.81	0.79	0.85
0.87	0.93	0.84	0.95	0.82	0.93	0.55	0.60	0.49	0.72
0.75	0.87	0.67	0.89	0.63	0.85	0.25	0.31	0.18	0.53
0.68	0.73	0.54	0.65	0.43	0.56	0.27	0.32	0.30	0.40
0.62	0.69	0.43	0.59	0.32	0.49	0.12	0.16	0.15	0.29
0.49	0.49	0.22	0.41	0.03	0.23	0.00	0.02	0.00	0.10

Based on these results, a question that one might ask is: Does the effectiveness of sequence alignment decrease or increase as IP accuracy increase? We would argue that effectiveness appears to increase, given a reasonable definition of effectiveness. We examined several plots constructed from the data in Table 2.6. We found a negative association between FP – IP and IP for Sp, CC, and Xsp, and no association for Sn and Xsn. These results are largely a consequence of the measures being bounded by a

maximum value of 1. Also, there is little improvement in Sn or Xsn when initial predictions are poor. Plots of proportional improvement $(FP - IP)/(1 - IP)$ versus IP reveal a different picture. In each plot there is a strong positive association. The slope is steepest for Sn, shallowest for XSn and XSp. These observations seem reasonable. We would expect a strong relationship between IP and proportional improvement in Sn. As IP increases, fewer exons are missed entirely and sequence alignment is more effective in improving sensitivity. Slopes for XSn and XSp are shallow because the denominators $(1 - IP)$ remain relatively large. Plots for Sp and CC are very similar. Values for AC are nearly the same as CC, and so are omitted from Table 2.6.

Table 2.7 Performance comparisons for Burset/Guigo set of 570 vertebrate genes with different protein database sizes.

Sequences	Sn	Sp	CC	AC	Xsn	Xsp
198742	0.972	0.967	0.961	0.961	0.809	0.854
158992	0.966	0.966	0.957	0.957	0.796	0.836
119245	0.954	0.959	0.944	0.945	0.780	0.817
99371	0.954	0.959	0.944	0.945	0.768	0.810
79497	0.944	0.950	0.932	0.933	0.753	0.787
66244	0.933	0.954	0.927	0.929	0.732	0.771
49685	0.934	0.948	0.925	0.927	0.722	0.760
37264	0.921	0.940	0.910	0.914	0.693	0.732
19874	0.922	0.934	0.907	0.910	0.687	0.712
0	0.929	0.928	0.908	0.911	0.758	0.788

Several authors (Snyder and Stormo, 1995; Burset and Guigo, 1996) have commented on a serious difficulty in comparing programs that incorporate sequence alignment as their performance depends on the database being searched. We investigated

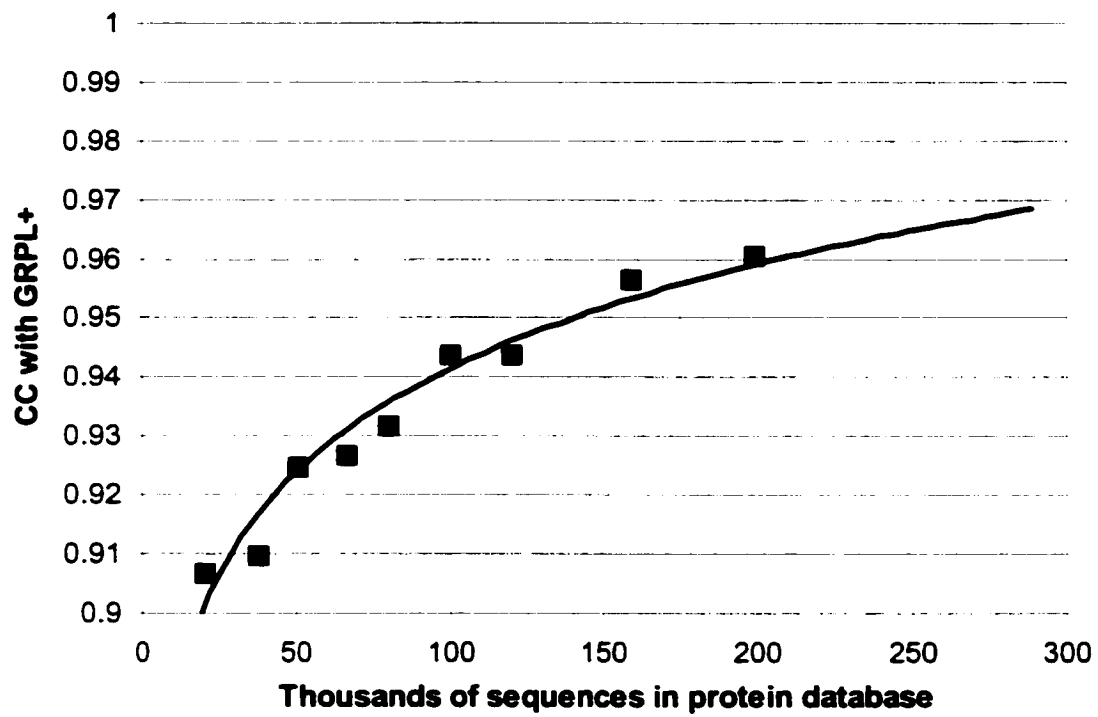


Figure 2.3 Performance of GRPL+ using the Burset/Guigo set of 570 vertebrate genes versus differing sizes of the OWL protein database.

the extent and nature of this dependency by applying GRPL+ using randomly selected subsets of varying size from the OWL database. We found an approximately linear relationship between each performance measure (Sn, Sp, Sq, CC, AC, XSn, XSp) and the logarithm of the subset size (see Table 2.7). Figure 2.3 displays a plot of CC versus subset size, with a logarithmic curve fitted to the points. We also found that, for measures at the nucleotide level, the GRPL results were improved by sequence alignment when only a small subset of the database was used. For measures at the exon level, however, a subset of 100,000 sequences (half of the database) was required before an improvement was seen. XSn and XSp were made worse by sequence alignment when smaller subsets were used. To further test effectiveness of sequence alignment, we predicted coding regions by two defective GRPL versions and one randomly generated prediction along with the full version of GRPL and then applied sequence alignment on top of these predictions. The results of this experiment are illustrated in Figure 2.4. We found that the performance of GRPL+ is somewhat related to the accuracy prediction levels of GRPL. The more accurate of the initial GRPL prediction, the smaller the effect of sequence alignment (GRPL+) on improving the prediction. This decreasing margin of improvement is expected, especially as the accuracy at the initial predictions approaches 100%.

2.4 Conclusion

The use of reference point logistic (RPL) classification and regression, both alone and in combination with other techniques, represents a novel approach to functional site identification and gene prediction. Comparisons with other methods indicate that GRPL

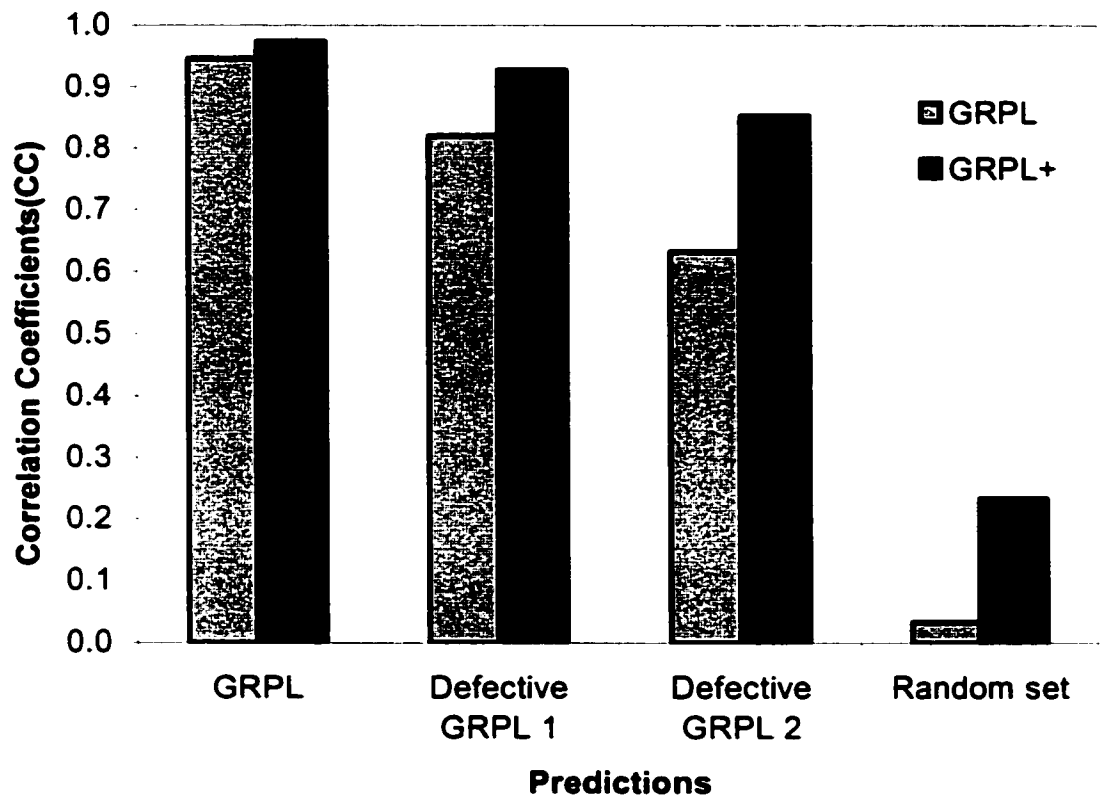


Figure 2.4 The performance of GRPL+ for 570 vertebrate genes compared to the same predictions for GRPL. CC = 1 means a perfect prediction.

can identify 5' and 3' splice sites with greater accuracy than other methods (Rogozin and Milanesi, 1997). Furthermore, by combining GRPL classification with more established approaches (dynamic programming, Generalized Hidden Markov Models, and database sequence alignment) we have shown that it is possible to match (using GRPL) or exceed (using GRPL+) the performance of many of the best gene prediction (Burset and Guigo, 1996; Burge and Karlin, 1997). Importantly, the exceptional performance of GRPL is not compromised by computational speed. Indeed, we estimate that GRPL is typically five to ten times faster than other high-performing methods. Additionally, GRPL has been adapted to deal with partial, single, and multi-gene sequences from a wide range of eukaryotic organisms, including vertebrates, invertebrates, and plants. This combination of speed, accuracy, and versatility should make GRPL (and GRPL+) a useful tool for analyzing gene structure in large-scale sequencing projects. Burset and Guigo (1996) previously demonstrated that the inclusion of sequence alignment information in exon predictions can improve the accuracy of the results. We have both confirmed this and have helped to rationalize these results in this chapter. Given the database trends illustrated in Figure 2.3, we believe that as the database continues to expand that may allow for close to perfect gene predictions. No doubt these methods and results should be of some interest to those working on deciphering the first draft of the human genome.

2.5 Availability

An academic implementation of GRPL and GRPL+ compiled for SUN workstations (Solaris 2.5 or higher), is available on <http://redpoll.pharmacy.ualberta.ca/download>. The training and test sets used in this work,

together with supplementary material, can be obtained at the same location. A commercial implementation is available as a component of GeneTool (*BioTools Inc.*, <http://biotools.com>).

Chapter 3

RefDB: A Database of Uniformly Referenced Protein Chemical Shift Assignments Derived from the BioMagResBank

3.1 Introduction

Chemical shifts are perhaps the most precisely measurable but the least accurately measured parameters in NMR spectroscopy. This curious state of affairs has arisen because, unlike most spectroscopic measurements, chemical shifts are relative. As such, chemical shifts are prone to numerous kinds of reporting and measurement errors. The problem with chemical shift measurement is particularly acute in biomolecular NMR. Indeed, the large number of chemical shifts that must be measured (hundreds to thousands), the variety of chemical shifts (^1H , ^{13}C , ^{15}N , ^{31}P), and the incredible range of solvent conditions (pH, temperature, salts, organic solvent mixtures) – all contribute to the problem. A further complication has been the historic reliance on many different chemical shift standards or chemical shift measurement protocols – many of which are now obsolete or widely considered to be irreproducible. The problems with chemical shift standardization have been discussed at length in a number of recent articles (Wishart and Sykes, 1994c; Iwadate et al., 1999; Cornilescu et al., 1999; Wishart and Case, 2001) and several suggestions or widely-agreed upon standards have been advocated (Wishart et al., 1995; Wishart and Sykes, 1994c; Maurer and Kalbitzer, 1996; Markley et al., 1998).

A key point raised by these authors has been the fact that biomolecular chemical shifts, in particular, contain a tremendously rich source of structural and dynamic

information. However, the structural and dynamic information contained in chemical shifts is subtle and, consequently, inaccurate or incorrectly referenced chemical shift measurements can easily blur or distort an exquisitely detailed picture of a biomolecule.

The BioMagResBank (Seavey et al., 1991) was established in 1991 to help address some of the problems and inconsistencies in biomolecular chemical shift reporting. Over the past 10 years the BMRB has given biomolecular NMR spectroscopists a superb opportunity to systematically assemble, compare and interpret chemical shifts. It has been through the BMRB, for instance, that a number of important chemical shift trends have been identified (Spera and Bax, 1991; Wishart et al., 1991; Wishart et al., 1992; Metzler et al., 1993; Gronenborn and Clore, 1994) and a variety of chemical shift theories or prediction/assignment schemes have been refined (Osapay and Case, 1994; Wishart and Nip, 1998; Beger and Bolton, 1997; Le and Oldfield, 1994). Throughout its 10-year history the BioMagResBank has served as a superb historical archive as it has meticulously recorded the ever-changing trends in chemical shift measurement and reporting. Because the BMRB is an archival database (it accepts "as-is" data directly from depositors) it depends crucially on the integrity and accuracy of its depositors. However, given the nature of chemical shift assignments and the variability of chemical shift reporting, it has been difficult to develop a rigorous set of protocols to validate the chemical shifts being deposited into the BMRB. As a result the BMRB likely contains a number of chemical shift assignments which have been improperly referenced or incorrectly assigned (Williamson et al., 1995; Iwadate et al., 1999; Wishart and Case, 2001). Indeed, a preliminary survey conducted in 2000 suggested that up to

20% of ^{13}C shifts and 30% of ^{15}N shifts are improperly referenced (Wishart and Case, 2001). This result is of some concern and it leads immediately to a number of important questions: What is the true magnitude of these referencing problems? What nuclei are most frequently or significantly affected? Do these affect the chemical shift trends or theories that have been developed from BMRB data? Can a corrected set of shifts be assembled? Can a chemical shift validation suite or protocol be developed for the BMRB?

Here we wish to report on the development of a set of software tools and a complementary chemical shift database (RefDB) containing a subset of BMRB chemical shifts that have been properly re-referenced according to the IUPAC/IUB conventions (Wishart et al., 1995; Markley et al., 1998). We also demonstrate how these analysis tools can be used not only to correctly reference chemical shifts, but to identify potential mis-assignments, to flag typographical errors, to detect spectral folding problems and zero-in on the location of potential structural differences or structure refinement errors. We also show how the RefDB shifts can be used to generate a more refined set of secondary shifts for all 20 amino acids, a tabulation which may be of some use in secondary structure analysis and empirical chemical shift calculations.

3.2 Materials and Methods

RefDB was prepared using a combination of three different computer programs. The first program (SHIFTX) calculates backbone ^1H , ^{13}C and ^{15}N chemical shifts from protein 3D coordinate data. The second program (SHIFTCOR) compares the calculated shifts with the observed shifts, evaluates any statistically significant differences and

performs the necessary chemical shift corrections. The third program (UPDATE) automatically retrieves newly deposited BMRB data along with any corresponding PDB data. UPDATE also directs the data to SHIFTCOR and appends the "corrected" chemical shift file to the RefDB database. A more detailed description of each program follows:

SHIFTX

SHIFTX uses a semi-empirical approach to calculate ^1H , ^{13}C and ^{15}N protein chemical shifts. The program employs both a combination of empirically derived chemical shift hypersurfaces (Spera and Bax, 1991; Le and Oldfield, 1994; Wishart and Nip, 1998) and classically calculated ring-current, electric field, nearest neighbor and hydrogen bond effects (Wagner, 1983; Wishart et al., 1991; Osapay and Case, 1991). The hypersurfaces, which relate ^1H , ^{13}C and ^{15}N chemical shifts to backbone dihedral angles, were derived from the chemical shift assignments of more than two dozen fully assigned and highly resolved ($< 1.8 \text{ \AA}$) X-ray structures in a manner similar to Iwadate et al. (1999). Ring current effects were calculated using the method of Haigh and Mallion (1979), whereas electric field and hydrogen bonding effects were calculated using methods similar to Osapay and Case (1991) and Wagner (1983). Nearest neighbor effects and local side chain effects were derived through a specialized data-mining program and incorporated into SHIFTX as empirical correction factors. Nucleus-specific constants were calculated for ring-current, electric field, nearest neighbor and hydrogen bond effects. The performance of SHIFTX was evaluated both on a training set (20 proteins) and a test set (10 proteins) each of which had high resolution ($< 1.80 \text{ \AA}$) X-ray structures with uniformly referenced chemical shifts. Overall, the program was able to

attain a correlation coefficient (r) between observed and calculated shifts of 0.905 ($^1\text{H}\alpha$), 0.973 ($^{13}\text{C}\alpha$), 0.996 ($^{13}\text{C}\beta$), 0.860 (^{13}CO), 0.891 (^{15}N) and 0.748 (^1HN). The RMS error was 0.25, 1.3, 1.2, 1.3, 3.4, 0.3 ppm for $^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN shifts, respectively. Relative to previously published shift prediction programs (Iwadate et al., 1999; Xu and Case, 2001; Osapay and Case, 1994) SHIFTX is uniquely able to calculate all measurable backbone chemical shifts (including ^{13}CO and ^{15}N shifts) with a very high degree of accuracy and precision. SHIFTX reads standard PDB-formatted files and outputs the predicted chemical shifts in a simple tabular form (BMRB or SHIFTY (Wishart et al., 1997) format). More complete details regarding the performance and structure of SHIFTX will be forthcoming shortly (Neal and Wishart, manuscript in preparation).

SHIFTCOR

SHIFTCOR is an automated shift correction program that uses statistical methods to compare and correct SHIFTX-predicted shifts relative to an input set of observed chemical shifts. SHIFTCOR uses several simple statistical approaches and pre-determined cutoff values to identify and correct potential referencing, assignment and typographical errors. The standard input for the SHIFTCOR program is a set of observed chemical shifts (BMRB or SHIFTY format) and a corresponding PDB file. SHIFTCOR identifies potential chemical shift referencing problems by comparing the difference between the average value of each set ($^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN) of observed and predicted chemical shifts. The average observed shifts are calculated after excluding potential mis-assignments or typographical errors to ensure these extreme outliers do not

bias the calculation. Potential mis-assignments are initially identified by looking for predicted chemical shifts that differ from their corresponding observed chemical shifts by approximately 4 standard deviations (i.e. 4x the RMS error expected for SHIFTX predicted shifts). Specifically the maximal cutoff differences were 0.7, 5.0, 5.0, 5.0, 10.0 and 2.0 ppm for $^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN shifts respectively. These values were determined after an extensive series of cut-off selection trials and later rounded up or down for ease of recall. Notice that the precise cut-off value differs slightly for each nucleus due to a combination of factors. When SHIFTCOR is run, it creates two files, one contains the chemical shift analyses (including lists of potential mis-assignments, estimates of the referencing error, correlation coefficients, RMSD values) and the other contains the corrected BMRB formatted chemical shift file (see Figure 3.1 for an example). Note that SHIFTCOR is not capable of detecting or classifying typographical errors (missing or added digits), switched assignments (i.e. Ser for Thr) or other anomalies. These were identified manually (after initially being identified as mis-assignments) and the corrections included in the current version of RefDB.

UPDATE

UPDATE is a database updating program designed to automatically process newly deposited protein chemical shift data in the BioMagResBank and store the results in the RefDB database. It can be divided into five steps (Fig. 3.2). Firstly, UPDATE uses standard web query protocols to identify and download newly deposited chemical shift data in the BioMagResBank. Second, after downloading the BMRB file, UPDATE reads the file description keywords to identify if the file has a corresponding PDB

The following residues have a HA chemical shift difference (obs-pred) greater than 0.7ppm:

#NUM	AA	CS	Observed	Predicted
# 17	Q	HA	4.06	5.40
# 76	F	HA	4.42	5.49
# 77	G	HA	5.20	4.02
# 78	Q	HA	4.24	5.07
# 98	L	HA	4.19	2.71
#102	A	HA	4.56	3.57
#104	K	HA	4.71	3.06
#105	F	HA	4.87	3.37
#107	E	HA	5.26	3.59
#109	K	HA	5.25	3.94

The following residues have a CB chemical shift difference (obs-pred) greater than 5.0ppm:

#NUM	AA	CS	Observed	Predicted
# 99	S	CB	56.56	62.37

The following residues have an N chemical shift difference (obs-pred) greater than 10.0ppm:

#NUM	AA	CS	Observed	Predicted
# 6	Q	N	121.01	131.52

The average CS difference between predicted and observed:
(Add these values to the corresponding observed chemical shifts.)

HA	CA	CB	CO	N	HN
0.06	0.17	0.30	0.06	-1.56	0.11

The Correlation Coefficient between predicted and observed:

HA	CA	CB	CO	N	HN
0.511	0.969	0.996	0.822	0.793	0.693

The RMSD between predicted and observed:

HA	CA	CB	CO	N	HN
0.381	0.654	0.947	0.673	2.240	0.336

Figure 3.1 SHIFTCOR output for bmr4766.str.

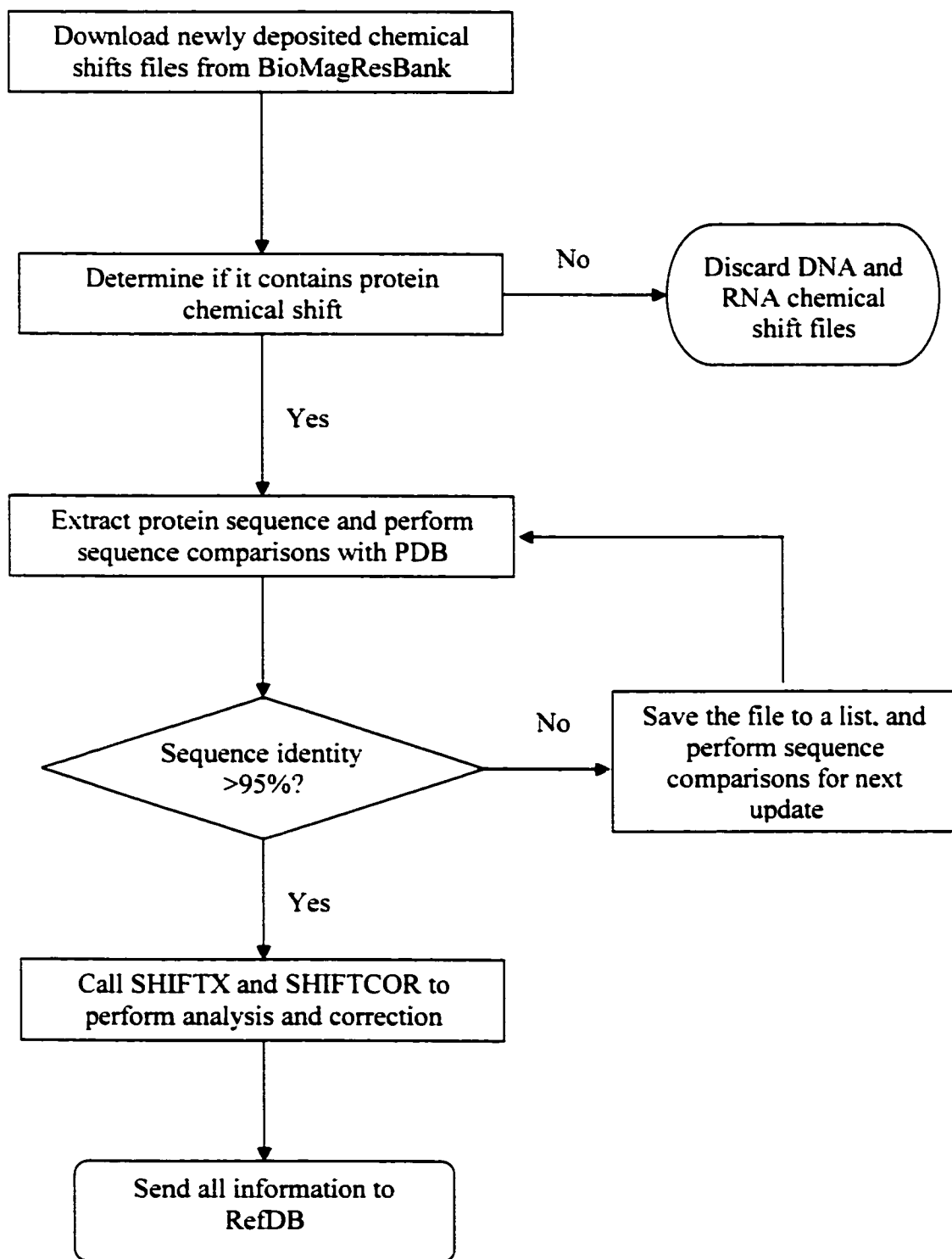


Figure 3.2 An outline of the UPDATE algorithm.

accession number. If the PDB code is found, UPDATE extracts the sequence from the BMRB file and uses a second web-based query to conduct a BLAST sequence search against the PDB. At least 95% sequence identity between the two sequences is required to identify a “matching” PDB file. If a single match is found, it is downloaded and processed. If more than one PDB file is found, the 3D coordinate file that is most highly resolved is selected. X-ray structures are given precedence over NMR structures because of their intrinsically higher resolution (Vriend, 1990; Laskowski et al., 1993). If the PDB file contains more than one structure (as is the case with many NMR data sets) UPDATE selects just one of the structures for processing. If the X-ray/NMR structure differs in length from the reported assignments, only those residues with 3D coordinates will have their chemical shifts calculated and adjusted. This can lead to an apparent “shortening” of the assignment list. After the appropriate PDB file has been selected and automatically downloaded, SHIFTX and SHIFTCOR are then called to perform their respective calculations and corrections. UPDATE then appends these corrected data files, along with the corresponding 3D coordinates to the RefDB database.

RefDB

Currently RefDB contains nearly 300 sets of corrected protein chemical shifts. All of the original chemical shift sets were obtained from the BioMagResBank. Each polypeptide in RefDB is required to contain at least 25 residues and to have an X-ray or NMR structure deposited in the PDB with backbone and side chain coordinates. RefDB does not include proteins dissolved in urea, TFE, DMSO or other organic solvents since

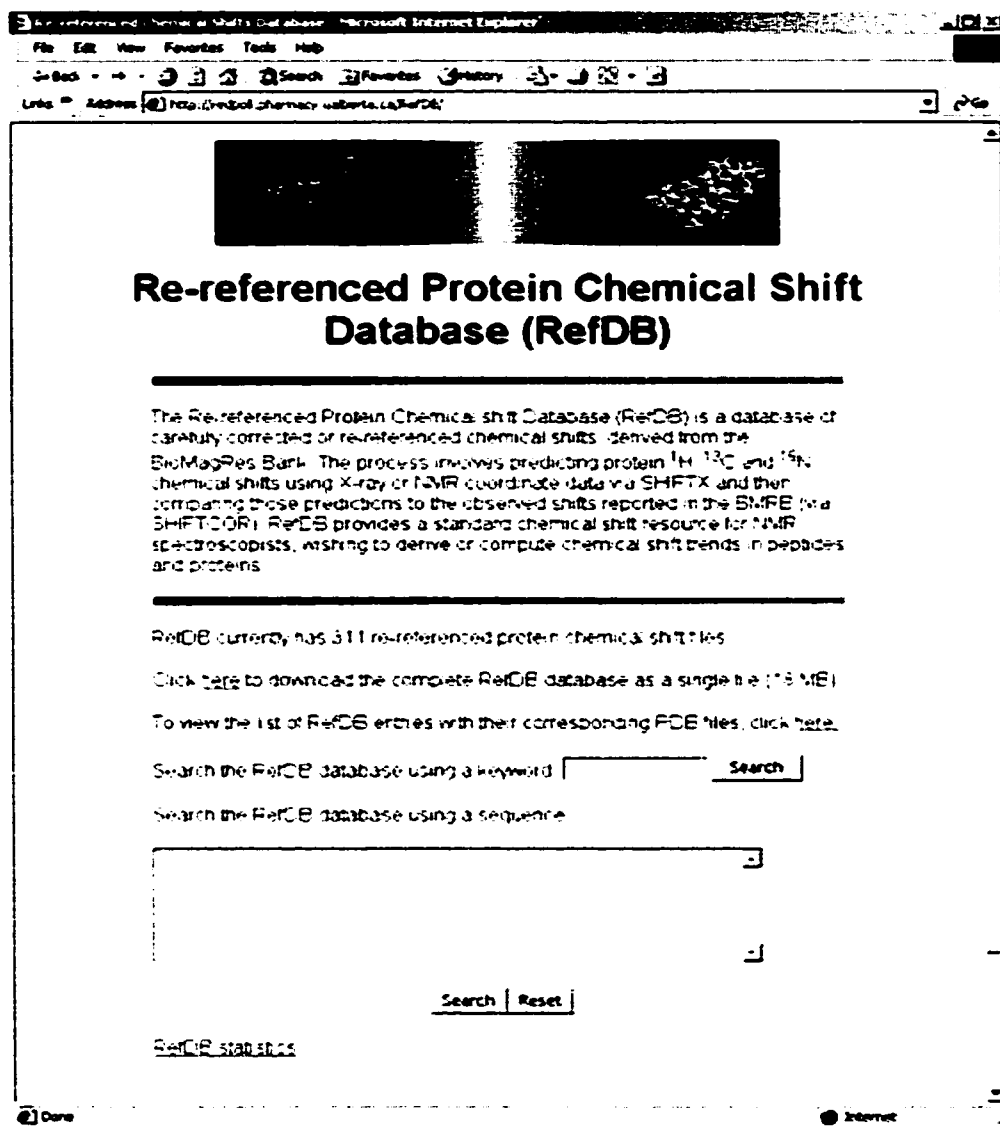


Figure 3.3 The Interface of RefDB database server.

these solvents can affect the chemical shift referencing in unpredictable ways (Wishart et al., 1995; Wishart and Nip, 1998). Furthermore, polypeptides dissolved in these solvents differ substantially from their native (X-ray) or reference structure. RefDB exists as both a single flat-file (~18 Megabytes) for convenient downloading, and as a web-enabled, queryable database. The RefDB web server is located at <http://redpoll.pharmacy.ualberta.ca> (Figure 3.3). The web version of RefDB uses a formatted table to list the name of the original BMRB file (hyperlinked to the BMRB site), the name of the corrected or adjusted shift file (hyperlinked to the shift list), the full name of the protein and the PDB accession number of the corresponding 3D structure (hyperlinked to the PDB). The web version of RefDB also supports a local BLAST sequence search (Altschul et al., 1997) as well as a fast boolean keyword query system supported by GLIMPSE (Manber and Wu, 1994; Manber and Bigot, 1998). This allows users to search RefDB via the sequence, partial sequence, protein name, author name, accession number, chemical shift or any other keyword or combination of keywords. All corrected protein chemical shift files archived in RefDB adhere to the BMRB star format, with the SHIFTCOR analysis placed at the top of each file as a set of comments. Individual files can be downloaded separately via the web. RefDB is updated weekly via the UPDATE program.

3.3 Results and Discussion

At the time of this writing, RefDB consists of 263 different proteins out of a total of ~400 fully (>80 % complete) assigned, non-redundant proteins in the BioMagResBank.

Of these 263 proteins, 41 contain only ^1H assignments, 43 have ^{15}N and ^1H assignments and 178 proteins have ^1H , ^{13}C and ^{15}N assignments. Of those proteins with reported ^{13}C assignments, 97% have at least $^{13}\text{C}\alpha$ shift assignments, 85% have both $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ shift assignments while just 55% have ^{13}CO shift assignments. A total of 121 proteins have at least one corresponding X-ray structure while 142 have only NMR derived structures. The smallest protein in RefDB is 25 residues (PDB 1HVW; bmr4937) and the largest is 370 residues (PDB 1ANF; bmr4354). There are a total of 150,057 corrected and re-referenced chemical shift assignments in the RefDB database. Statistics concerning the size and composition of RefDB are updated weekly and posted at the RefDB web site.

Of particular interest for this study was a precise determination of the magnitude and extent of chemical shift errors or problems in the BioMagResBank. Based on previous experience, we identified five types of potentially classifiable chemical shift "errors" including: 1) referencing errors; 2) typographical errors; 3) assignment switches; 4) mis-assignments; and 5) structural discrepancies. With the possible exception of referencing errors, the latter four types of chemical shift errors have to be inferred on the basis of manual inspection or "reconstruction" of the assignment process. Some of these errors are easily identified, while others are far more subtle. For instance, the addition or deletion of digits or decimal points (84.0 vs. 8.40 for a ^1HN shift) is an obvious typographical error, whereas the exchange of two digits (8.34 vs. 8.43 for a ^1HN shift) is almost undetectable. As a general rule, if we couldn't classify a chemical shift anomaly

as either a typographical error, an assignment switch/exchange or a mis-assignment, we would attribute it to a structural discrepancy (solution vs. solid state).

Referencing Errors

Referencing errors or referencing adjustments are systematic errors arising from the improper referencing of ^1H , ^{13}C or ^{15}N chemical shifts. Most NMR spectroscopists are quite diligent in their chemical shift referencing protocols. However, even the most careful worker can make mistakes. These mistakes may arise from 1) incorrect instrument settings; 2) data processing errors; 3) sample preparation or decay; 4) failure to account for isotopic shifts; 5) failure to adhere or failure to understand IUPAC/IUB referencing protocols; or 6) use of obsolete referencing standards (TMS, NH_4Cl , H_2O). These kinds of systematic errors are of considerable concern in biomolecular NMR because they can affect nearly every chemical shift assignment. Furthermore, they can often be sufficiently large to make almost all secondary shifts undetectable or misleading (Wishart et al., 1995). What is most frustrating is that these types of chemical shift errors, particularly for ^{13}C and ^{15}N nuclei, have often been exceedingly difficult to identify.

In this study we investigated the occurrence of referencing errors for each type of nucleus (^1H , ^{13}C and ^{15}N) separately. To validate our methods for detecting referencing errors, we conducted exhaustive comparisons to a number (>30) of fully (^1H , ^{13}C , ^{15}N) assigned proteins and the reported referencing procedures provided in either the BMRB or the associated literature. Our results showed that the method was able to consistently detect and correctly quantify the magnitude of the systematic error based on published shift correction tables (Wishart and Case, 2001). Because of these manual comparisons,

we are confident that the predicted chemical shift adjustments calculated by SHIFTCOR and presented in RefDB are accurate.

The first set of shifts we analyzed in RefDB was the ^1H shifts. Because almost all ^1H chemical shifts are determined using an internal primary reference (DSS, TSP) or a well characterized secondary reference (HDO) one would not expect to find any significant ^1H shift referencing errors. Indeed the data in RefDB bear this out as we found no significant referencing errors among ~255 sets of ^1H assignments. The largest difference between any set of observed and predicted $^1\text{H}\alpha$ shifts was 0.27 ppm (bmr4952) with the vast majority of $\text{H}\alpha$ referencing errors being less than 0.10 ppm. On the other hand, because of the long-standing confusion over how to indirectly (or directly) reference ^{13}C or ^{15}N shifts, we found there were many more significant problems with these shifts. For instance, 60/212 (28.3%) of proteins with ^{15}N assignments required reference adjustments (up or down) of more than 1 ppm. Furthermore, 45/178 (25.3%), 51/161 (31.7%) and 23/105 (21.9%) of proteins in RefDB required reference adjustments of more than 0.5 ppm for their reported $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$ and ^{13}CO assignments, respectively.

Table 3.1 The number of proteins and their associated ranges of referencing errors for ^{13}C and ^{15}N chemical shifts.

Referencing error (ppm)	$^{13}\text{C}\alpha$	$^{13}\text{C}\beta$	^{13}CO	^{15}N
0.5~1.0	16	23	10	(54)
1.0~1.5	12	11	4	31
1.5~2.0	9	9	4	17
>2.0	8	8	5	12
Total	45	51	23	60

Although there are many proteins that required chemical shift adjustments, most of these re-referencing changes fell into the range of 0.5-1.0 ppm for ^{13}C shifts and 1.0-1.5 ppm for ^{15}N shifts (Table 3.1). As might be expected, the relative frequency of these $^{13}\text{C}/^{15}\text{N}$ chemical shift referencing errors falls off exponentially relative to their magnitude. The largest referencing adjustment required for ^{13}C shifts was more than 3 ppm (bmr4431), whereas the largest ^{15}N chemical shift adjustment was 4.1 ppm (bmr4127). Among those proteins identified as requiring significant adjustments were seven proteins which were fully deuterated (bmr4354, bmr4775, bmr4836, bmr4936, bmr4986, bmr4987, vmr5161). Since the ^{13}C and ^{15}N chemical shifts of deuterated proteins are shifted upfield (0.43 ppm for $^{13}\text{C}\alpha$, 0.82 ppm for $\text{C}\beta$, and 0.23 ppm for ^{15}N) relative to those expected for a fully protonated sample (Gardner et al., 1997; Bjorndahl et al., 2001), these chemical shift differences should not be classified as referencing errors. Indeed, their reported chemical shift displacement suggests that all seven samples were correctly referenced according to IUPAC conventions.

Figure 3.4 and Figure 3.5 plots the frequency of chemical shift referencing errors for $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO and ^{15}N assignments versus the year of reporting/deposition. As can be seen from these graphs, heteronuclear chemical shift referencing problems were especially widespread prior to 1994. After 1995 there appears to be a significant improvement, indicating a stricter adherence to IUPAC/IUB ^{13}C and ^{15}N chemical shift referencing recommendations (Wishart and Sykes, 1994c; Wishart et al., 1995; Markley et al., 1998). Interestingly, more than five years after the recommendations were first made, we still see that approximately 20% of newly deposited protein chemical shifts are

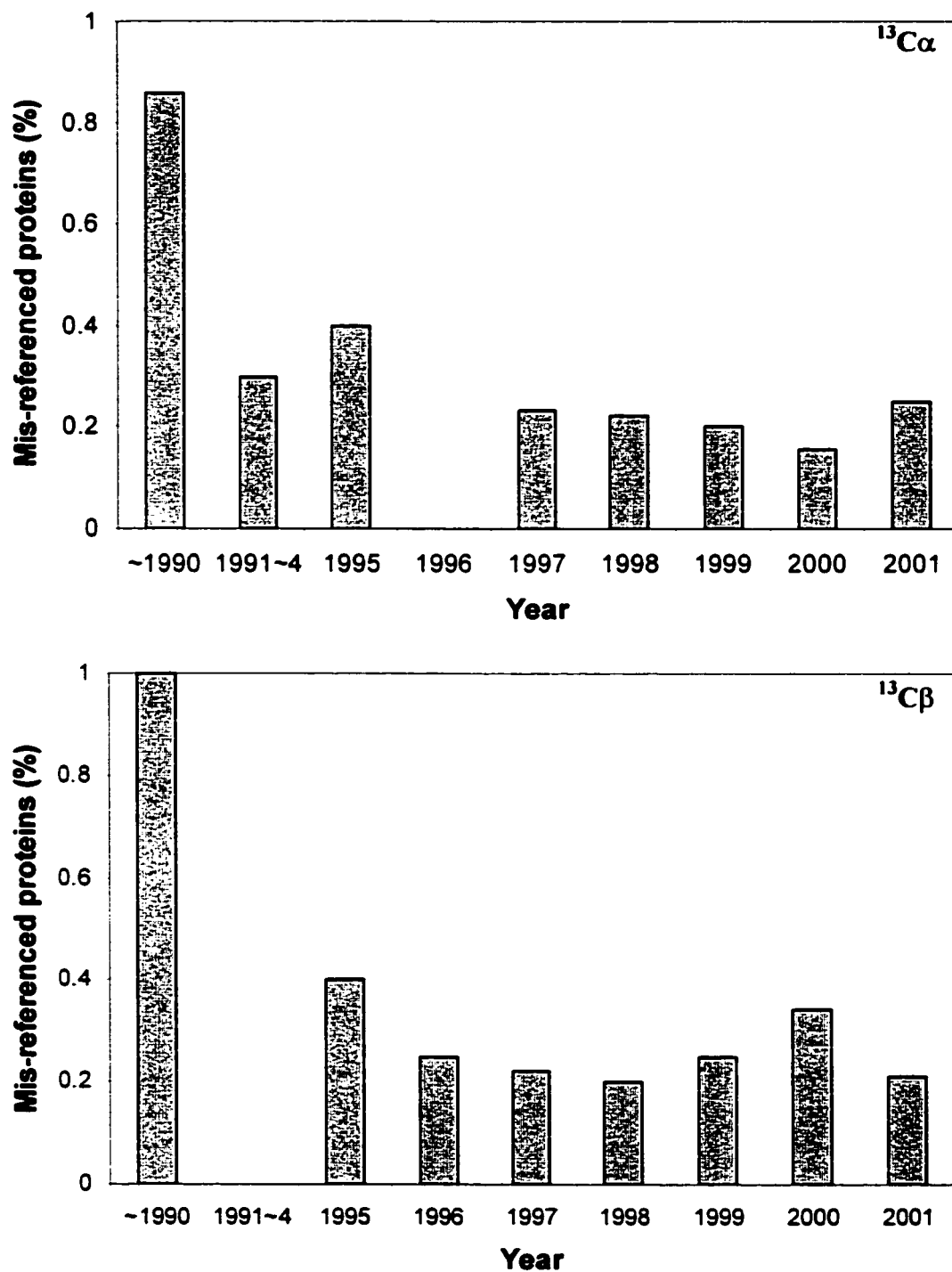


Figure 3.4 The percentage of referencing errors for $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ shifts versus year of deposition/submission.

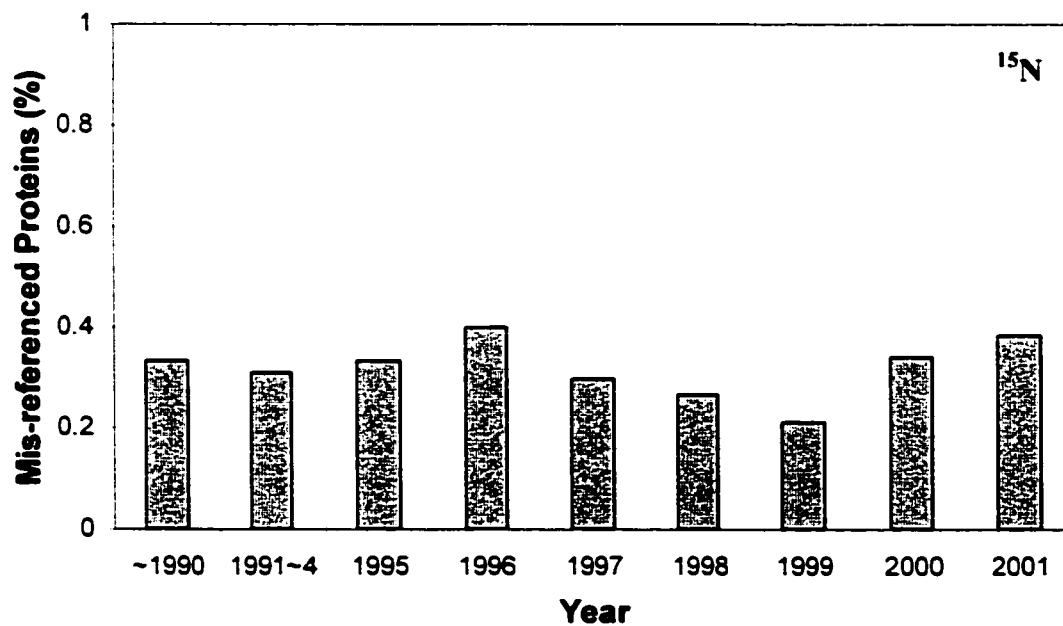
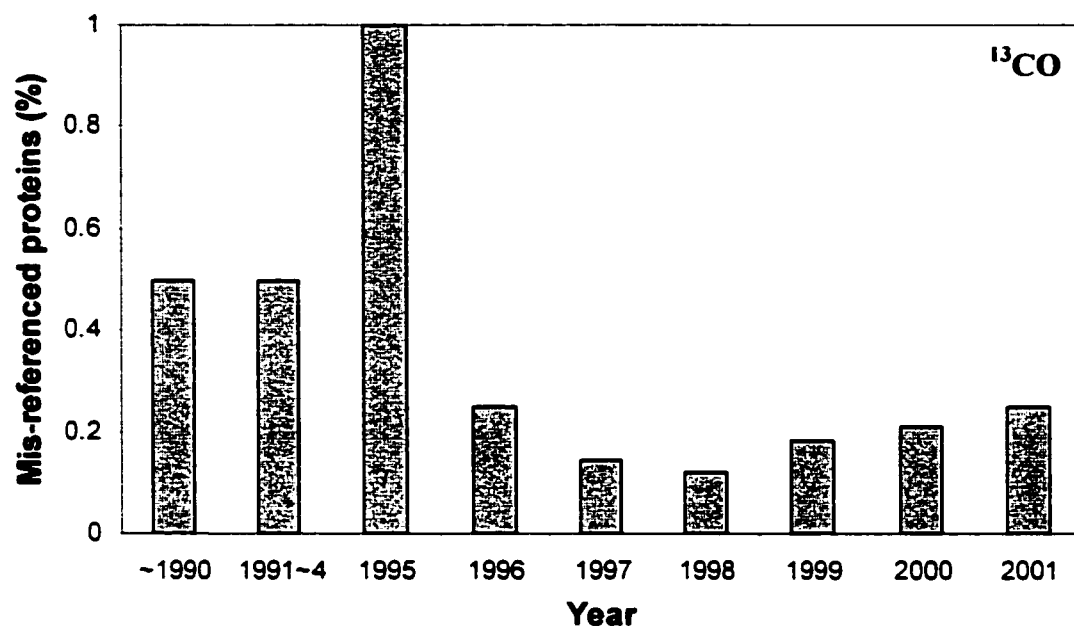


Figure 3.5 The percentage of referencing errors for ^{13}CO and ^{15}N shifts versus year of deposition/submission.

improperly referenced. This suggests that chemical shift referencing is still problematic for a significant number of individuals in the biomolecular NMR community.

Outside of improved education, improved lab practices and stricter rules about adherence to IUPAC recommendations, it may be that the best approach to dealing with this problem is to use computer programs such as SHIFTCOR as an integral part of the data checking/validation process prior to submitting or accepting data at the BMRB. Similar data checking and validation procedures for PDB coordinate submission are either freely available to structural biologists (Lin et al., 2000) or are already in place at the RCSB (Berman et al., 2000). Indeed, the development of data validation and data checking programs have become a major thrust for just about every major biological or bioinformatic database (Lin et al., 2000).

Mis-Assignments, Typo's and Other Errors

While our principle concern was to develop software tools and methods to identify and fix chemical shift referencing errors, we found that other chemical shift

Table 3.2 Number and type of assignment-related errors (263 proteins).

Type of error	¹³ C α	¹³ C β	¹³ CO	¹⁵ N	¹ HN	¹ H α
Mis-assignment	31	27	194	30	5	N/A
Labelling/Typographical	2	1	N/A	1	N/A	2
Struct difference	N/A	N/A	N/A	4	5	171
Switches	2	2	N/A	2	2	N/A
Switch/Typographical	2	5	N/A	N/A	N/A	3
All categories	37	35	194	37	12	176
Total assignments	23205	14210	13051	23356	30068	29140

errors could also be detected. Indeed, as Williamson et al. (1995) has already pointed out, accurate, structure-based chemical shift calculations can be used quite effectively to identify ^1H assignment errors. Unlike systematic referencing errors, these “random” errors are not easily classified (or identified) without manual inspection or some prior knowledge about the nuances of the NMR assignment process. Similarly, the correction of these errors also requires manual intervention.

Table 3.2 provides a summary of the number and type of assignment-related errors that were manually identified. It must be emphasized that these are “probable” errors as we cannot confirm their origin or cause without access to the raw experimental data. In all likelihood this is an underestimate of the true number of errors in the data set. As seen in this table, we identified 6 typographical errors in 5 different proteins (bmr4087, bmr4115, bmr4126, bmr4726, bmr4894). We also found four instances of $^{13}\text{C}\alpha/^{13}\text{C}\beta$ switches (bmr4068 and bmr4050) – two for threonine and two for serine. These assignment switches are quite understandable in light of the unusual downfield $^{13}\text{C}\beta$ shifts for these amino acids and their proximity to $^{13}\text{C}\alpha$ values. We also identified 2 assignment switches for ^1HN and 2 for ^{15}N (bmr4082 and bmr4344). Another 10 assignments were identified as either switches or typographical errors but could not be definitively classified in one or the other category. For those ^{13}C and ^{15}N shifts that differed by more than 5-6 standard deviations from the predicted values, but which fell within the allowed range of ^{13}C or ^{15}N shifts (regardless of amino acid type), we classified as “mis-assigned”. Clearly, some of these resonances may be correctly assigned and that their substantive differences arose from structural effects or our

imperfect understanding of chemical shift principles. Nevertheless, their level of abundance (~0.5 %) and past experience with other NMR assignment data sets suggests that this is well within the range of expected mis-assignments.

On the other hand, for many of the ^1H shifts it was essentially impossible to determine whether substantial shift discrepancies arose from mis-assignments or from structural differences. Consequently, we chose to err on the side of caution and ascribed these extreme outliers to probable structural differences (solid vs. liquid state, imperfect refinement, N or C terminal changes, etc.) as opposed to mistaken assignments. An example of the kind of issues one might find when analyzing ^1H shifts is found in bmr4766 (Fig. 3.1). As can be seen in this example, there are two near-contiguous regions exhibiting larger-than-expected deviations in their $^1\text{H}_\alpha$ chemical shifts. The chemical shift differences seen for residues 76-78, likely arises from structural differences between the solution and crystal state. Specifically, the ring of Phe 76 is probably much closer to the backbone in the crystal structure than in solution, thereby leading to an upfield ring-current induced shift for nearby $^1\text{H}_\alpha$ nuclei. On the other hand, the chemical shift differences seen for residues 102-109, most assuredly arises from the fact that the protein structure solved by X-ray crystallography was shorter than the protein assigned by NMR. This C-terminal truncation in the X-ray structure likely lead to a real structural change (i.e. the loss of a helix) that is manifested in the substantially different $^1\text{H}_\alpha$ chemical shifts for this region. Given the different conditions and different samples used by X-ray crystallographers relative to NMR spectroscopists, these kind of small discrepancies were not uncommon, nor were they unexpected.

Perhaps the most dramatic example of an assignment error in RefDB was found for the $^{13}\text{C}\text{O}$ resonances in bmr4775. While displaying good overall correlations for $^1\text{H}\alpha$ (0.709), ^{15}N (0.698) and $^{13}\text{C}\alpha$ (0.947) shifts, we found the $^{13}\text{C}\text{O}$ shifts were strongly negatively correlated (-0.765)! As no other protein analyzed by SHIFTCOR had shown a negative correlation for any set of chemical shifts, we decided to investigate this situation further. On closer inspection it became obvious that the $^{13}\text{C}\text{O}$ spectrum for this protein must have been folded prior to its assignment (perhaps due to the use of incorrect offset pulses, a far too narrow sweep width or inappropriate data processing). Given the intrinsically narrow range of $^{13}\text{C}\text{O}$ shifts and the lack of any kind of characteristic “marker” shifts (such as those seen with glycine for ^{15}N and $^{13}\text{C}\alpha$) it is not difficult to understand how this kind of error could be made nor how it could go undetected.

Re-evaluating Secondary Chemical Shifts

While the primary purpose of this exercise was to identify, enumerate and correct chemical shift referencing and chemical shift assignment errors, we also wanted to demonstrate how this “corrected” data could be used in a more practical sense. One obvious application would be to use this data to improve upon the accuracy of chemical shift calculation routines (Iwadate et al., 1999; Osapay and Case, 1994; Wishart and Nip, 1998; Wishart and Neal, in preparation). A second application might be to improve upon secondary structure identification (Metzler et al., 1993; Wishart et al., 1992; Wishart et al., 1995) or in dihedral angle calculation (Cornilescu et al., 1999). A third application might be in developing more accurate or consistent methods for alignment-based chemical shift prediction (Wishart et al., 1997; Potts and Chazin et al., 1998).

Rather than attempt to address all three areas here, we decided to focus on re-evaluating the so-called secondary chemical shifts or secondary-structure induced shifts associated with ^1H , ^{13}C and ^{15}N nuclei. To generate this data set, corrected chemical shifts from RefDB were assembled for each residue type along with the experimentally observed secondary structure. Secondary structures were calculated directly from PDB files using program VADAR (Wishart et al., 1994d). Because it is based on objective measures of peptide geometry, VADAR provides a far more consistent assignment of secondary structure location than those made by individual crystallographers or NMR spectroscopists. The results of these calculations are shown in Tables 3.3 – 3.8 where we have calculated the average characteristic shifts for residues in helices, beta-strands and “coil” regions for $^1\text{H}\alpha$, ^1HN , $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO and ^{15}N nuclei. With 150,057 “corrected” assignments, this collection represents the largest and most complete set of shifts for which this kind of calculation has been done (Wishart et al., 1991; Wishart and Sykes, 1994; Wishart and Nip, 1998). Given that the smallest number of assignments for any one category was still 39 (^{13}CO assignments for tryptophans in helices), we can be quite confident about the statistics (mean, median, standard deviation, range, etc.) for these numbers.

Overall, these chemical shifts show a very good level of agreement relative to previously published sets (Wishart et al., 1991; Wishart and Sykes, 1994c; Wishart and Nip, 1998), with the possible exception of some of the less-abundant residues and/or nuclei (esp. tryptophan, methionine and histidine). Interestingly, with a much larger data

Table 3.3 Averaged $^{13}\text{C}\alpha$ chemical shift values categorized according to secondary structural assignment (total number of residues observed in parentheses).

Residue type	Coil		Helix		Beta strands	
Ala	52.77	(404)	54.63	(625)	51.46	(279)
Cys	56.87	(97)	60.25	(89)	56.02	(101)
Asp	54.14	(550)	56.29	(303)	53.91	(195)
Glu	56.85	(423)	58.99	(599)	55.58	(260)
Phe	57.77	(187)	60.56	(224)	56.69	(264)
Gly	45.40	(912)	46.54	(191)	45.18	(211)
His	56.00	(151)	58.62	(125)	55.20	(90)
Ile	61.35	(206)	64.35	(309)	59.92	(364)
Lys	56.65	(503)	58.70	(429)	55.44	(314)
Leu	55.12	(401)	57.24	(575)	53.93	(387)
Met	55.74	(113)	58.11	(192)	54.43	(85)
Asn	53.27	(435)	55.20	(193)	52.58	(180)
Pro	63.18	(451)	65.11	(103)	62.71	(149)
Gln	56.18	(250)	58.34	(307)	54.74	(144)
Arg	56.38	(313)	58.87	(330)	55.39	(210)
Ser	58.38	(531)	60.64	(257)	57.49	(299)
Thr	61.60	(399)	65.39	(251)	61.20	(406)
Val	61.98	(275)	65.96	(386)	60.78	(531)
Trp	58.13	(69)	59.91	(67)	56.63	(87)
Tyr	57.79	(183)	60.99	(162)	56.77	(239)
Total number of chemical shifts	6748		5630		4727	

Table 3.4 Averaged $^{13}\text{C}\beta$ chemical shift values categorized according to secondary structural assignment (total number of residues observed in parentheses).

Residue type	Coil		Helix		Beta strands	
Ala	19.10	(343)	18.16	(511)	21.10	(249)
Cys(ox)	40.44	(43)	39.52	(33)	43.19	(36)
Cys(red)	29.86	(43)	27.49	(40)	30.28	(35)
Asp	40.77	(449)	40.38	(247)	42.12	(162)
Glu	30.12	(350)	29.29	(468)	31.82	(225)
Phe	39.69	(153)	38.93	(187)	41.39	(219)
Gly	N/A	N/A	N/A	N/A	N/A	N/A
His	29.59	(125)	29.29	(101)	31.70	(74)
Ile	38.44	(180)	37.54	(253)	39.96	(317)
Lys	32.64	(408)	32.17	(351)	34.51	(259)
Leu	42.26	(323)	41.66	(456)	44.02	(340)
Met	32.87	(84)	32.01	(142)	34.58	(73)
Asn	38.53	(358)	38.35	(152)	39.97	(160)
Pro	32.07	(380)	31.31	(85)	32.22	(123)
Gln	29.17	(209)	28.30	(263)	31.48	(124)
Arg	30.68	(259)	29.95	(265)	32.26	(171)
Ser	63.94	(431)	63.06	(195)	65.10	(249)
Thr	69.89	(326)	68.86	(297)	70.73	(340)
Val	32.54	(226)	31.44	(318)	34.00	(444)
Trp	29.78	(55)	28.97	(53)	31.44	(74)
Tyr	38.84	(145)	38.21	(140)	41.05	(198)
Total number of chemical shifts	4890		4448		3872	

Table 3.5 Averaged ^{13}C O chemical shift values categorized according to secondary structural assignment (total number of residues observed in parentheses).

Residue type	Coil		Helix		Beta strands	
Ala	177.65	(242)	179.38	(362)	176.26	(146)
Cys	174.94	(64)	176.47	(55)	174.06	(56)
Asp	176.33	(359)	177.89	(180)	175.84	(122)
Glu	176.55	(255)	178.47	(345)	175.60	(160)
Phe	175.74	(99)	176.93	(148)	174.36	(173)
Gly	174.03	(511)	175.43	(122)	172.85	(109)
His	174.98	(72)	176.80	(60)	174.04	(58)
Ile	175.63	(114)	177.51	(183)	174.86	(219)
Lys	176.39	(304)	178.29	(234)	175.42	(187)
Leu	176.92	(244)	178.31	(335)	175.76	(230)
Met	175.61	(54)	178.08	(123)	174.82	(55)
Asn	175.12	(255)	176.71	(109)	174.51	(97)
Pro	176.80	(223)	178.45	(69)	176.19	(80)
Gln	175.98	(148)	178.05	(179)	174.88	(86)
Arg	176.29	(155)	178.05	(185)	175.24	(104)
Ser	174.45	(310)	175.69	(175)	173.76	(173)
Thr	174.71	(240)	176.13	(142)	173.94	(218)
Val	175.86	(138)	177.62	(224)	174.81	(322)
Trp	176.23	(43)	177.80	(39)	175.58	(43)
Tyr	175.37	(92)	177.04	(86)	174.74	(127)
Total number of chemical shifts	3853		3270		2728	

Table 3.6 Averaged ^{15}N amide chemical shift values categorized according to secondary structural assignment (total number of residues observed in parentheses).

Residue type	Coil		Helix		Beta strands	
Ala	122.93	(459)	121.20	(701)	124.51	(329)
Cys	118.69	(122)	117.91	(101)	121.39	(116)
Asp	119.62	(609)	118.97	(338)	122.20	(225)
Glu	120.19	(472)	118.81	(649)	121.73	(295)
Phe	119.99	(202)	118.85	(250)	121.01	(298)
Gly	109.19	(1013)	107.16	(198)	109.33	(228)
His	118.32	(156)	117.88	(139)	120.55	(101)
Ile	120.97	(241)	119.56	(330)	122.74	(419)
Lys	120.36	(544)	118.86	(490)	122.40	(348)
Leu	121.22	(457)	119.50	(651)	124.30	(425)
Met	119.27	(109)	117.73	(207)	121.53	(92)
Asn	117.96	(481)	117.13	(214)	121.39	(191)
Pro	N/A	N/A	N/A	N/A	N/A	N/A
Gln	119.38	(281)	117.97	(338)	120.95	(173)
Arg	120.64	(345)	118.55	(371)	121.95	(225)
Ser	115.77	(584)	114.67	(296)	116.80	(332)
Thr	113.43	(436)	114.49	(278)	116.90	(491)
Val	119.93	(322)	119.04	(415)	121.95	(593)
Trp	120.88	(82)	119.79	(70)	122.17	(104)
Tyr	119.10	(195)	118.90	(187)	121.59	(264)
Total number of chemical shifts	7018		6124		5214	

Table 3.7 Averaged $^1\text{H}\alpha$ chemical shift values categorized according to secondary structural assignment (total number of residues observed in parentheses).

Residue type	Coil		Helix		Beta strands	
Ala	4.26	(499)	4.05	(645)	4.80	(299)
Cys	4.71	(198)	4.19	(115)	5.12	(163)
Asp	4.58	(632)	4.47	(324)	4.87	(210)
Glu	4.30	(492)	4.03	(616)	4.76	(285)
Phe	4.60	(215)	4.18	(227)	4.99	(281)
Gly	3.95	(1040)	3.82	(195)	4.18	(206)
His	4.54	(170)	4.36	(138)	5.00	(101)
Ile	4.16	(273)	3.70	(327)	4.67	(410)
Lys	4.24	(603)	4.02	(472)	4.71	(355)
Leu	4.32	(486)	4.02	(619)	4.84	(420)
Met	4.40	(130)	4.07	(208)	4.89	(97)
Asn	4.66	(502)	4.51	(202)	5.01	(181)
Pro	4.37	(531)	4.23	(109)	4.58	(151)
Gln	4.27	(292)	4.00	(333)	4.83	(165)
Arg	4.29	(388)	3.99	(374)	4.73	(224)
Ser	4.47	(635)	4.28	(294)	4.90	(327)
Thr	4.44	(471)	4.03	(283)	4.81	(472)
Val	4.15	(363)	3.62	(396)	4.61	(576)
Trp	4.64	(89)	4.37	(67)	5.15	(99)
Tyr	4.54	(210)	4.10	(182)	5.07	(246)
Total number of chemical shifts	7953		6006		5181	

Table 3.8 Averaged ^1HN chemical shift values categorized according to secondary structural assignment (total number of residues observed in parentheses)

Residue type	Coil		Helix		Beta strands	
Ala	8.22	(546)	8.04	(740)	8.51	(341)
Cys	8.36	(202)	8.14	(129)	8.72	(167)
Asp	8.29	(673)	8.17	(348)	8.57	(231)
Glu	8.34	(530)	8.22	(673)	8.53	(309)
Phe	8.26	(231)	8.16	(257)	8.70	(308)
Gly	8.33	(1145)	8.21	(224)	8.37	(245)
His	8.16	(183)	8.03	(154)	8.62	(108)
Ile	8.02	(282)	8.02	(356)	8.67	(450)
Lys	8.22	(640)	7.94	(531)	8.52	(382)
Leu	8.08	(527)	8.05	(683)	8.67	(445)
Met	8.26	(121)	8.05	(215)	8.67	(104)
Asn	8.40	(545)	8.19	(236)	8.64	(209)
Pro	N/A	N/A	N/A	N/A	N/A	N/A
Gln	8.17	(303)	8.03	(362)	8.52	(181)
Arg	8.21	(408)	8.11	(403)	8.52	(254)
Ser	8.26	(675)	8.06	(323)	8.55	(349)
Thr	8.16	(508)	8.09	(299)	8.54	(523)
Val	8.08	(386)	8.05	(438)	8.64	(643)
Trp	8.06	(99)	8.19	(76)	8.63	(113)
Tyr	8.00	(220)	8.15	(200)	8.70	(282)
Total number of chemical shifts		7979		6526		5563

set and better chemical shift referencing, the upfield/downfield trends for helices and beta sheets are now much more obvious for ^{15}N , ^1HN and ^{13}CO resonances. These trends had likely been obscured in previous studies because of the “noise” arising from improperly referenced chemical shift assignments. Looking at the results obtained for each of the 20 different residues for any given nucleus, it is also obvious that there are certain residue-specific trends concerning the extent of the upfield/downfield shifts. These may reflect intrinsic structural limitations (restricted phi/psi or side chain chi angles) or a statistical proclivity to be located in less mobile (or more mobile) regions of a polypeptide.

As indicated earlier, these tables may be of some utility in predicting chemical shifts (Wishart and Nip, 1998), in assessing preliminary chemical shift assignments, in automating chemical shift assignments (Moseley and Montelione, 1999), in identifying secondary structure (Wishart et al., 1992; Metzler et al., 1993) or evaluating nearest neighbor effects (Schwarzinger et al., 2001)

RefDB as a New Model for Bioinformatic Databases

With the increasing movement of towards storing vast quantities of biological data on electronic databases, it is clear that data handling and data storage will become increasingly important for just about everyone in the life sciences. Given the difficulty associated with handling and assimilating so much data from so many sources, we believe that it will be important to develop new approaches for automatically handling and analyzing biological data. In our view, RefDB may serve as a useful model for a new generation of self-updating, self-correcting bioinformatic databases. Specifically RefDB

makes use of the fact that all of the data it needs can be retrieved from the web through automated data mining tools (web-bots or web-spiders), automatically checked and modified (through resident data validation/checking software) and automatically displayed or accessed (via a self-updating web interface and CGI scripts). In other words, unlike any other biological database we are aware of, RefDB was designed to function autonomously, without the need for human intervention or human data entry. While the removal of the “human factor” from the database side does have its occasional down-side (run-away processes, mix-ups due to unannounced data format changes), we have been operating and updating RefDB continuously for the better part of a year, without the need for any student annotators or dedicated staff members. Furthermore, the data in RefDB is never more than 1 week out of date and never subject to slow-downs due to staff turnover or holidays.

The concept of self-updating databases appears to be relatively new and yet given the abundance of web-based tools, it is something that can be relatively easily implemented. While it was not our intention to break new ground in bioinformatics concepts, it appears that the ideas behind RefDB could be generalized to a much wider variety of biological or chemical databases.

3.4 Conclusion

There can be little doubt that chemical shifts are playing an increasingly important role in biomolecular NMR. Not only are they the “mileposts” which map atomic structure to NMR detectable parameters, but they also provide a means for NMR spectroscopists to

share and exchange raw experimental data. With the observation that chemical shifts contain a considerable amount of useful structural information, the importance of chemical shifts in biomolecular NMR has grown even further (Wishart and Sykes, 1994c; Szilagyi, 1995; Case, 2000; Williamson and Asakura, 1997). However, much of the utility of chemical shifts, both for assignment and for structural purposes depends on their accuracy and reliability. Recently, this reliability has been called into question (Wishart and Case, 2001).

In this study we have demonstrated that a significant portion of ^{13}C and ^{15}N chemical shift assignments made prior to 1995 need to be re-referenced – in some cases by as much as 4 ppm. We have also demonstrated that, while NMR spectroscopists are increasingly adhering to IUPAC recommendations, at least 20% of newly deposited protein chemical shifts are still improperly referenced. Furthermore, it appears that approximately 1% of all reported assignments may also be mis-assigned. In an effort to help sort out these persistent chemical shift referencing problems and to assist with the identification of potential mis-assignments, we have developed a self-updating database (RefDB) and a set of computational tools (SHIFTX, SHIFTCOR, UPDATE). As shown here, these tools should help correct these problems and facilitate both chemical shift analysis and chemical shift referencing.

Specifically, we believe RefDB and its associated programs could serve as: 1) a suite of programs and a set of criteria with which to assess, annotate and correct new (or old) BMRB entries; 2) a suite of programs and set of criteria with which individuals can

assess and correct their own assignments and structures (during refinement, or prior to submission); 3) a resource to help test, refine and develop chemical shift prediction programs; 4) a resource with which to test, refine and develop methods to predict protein structural features (helix caps, beta turns) from chemical shift data; and 5) a resource from which accurate chemical shift dependent patterns (secondary shifts, periodicity in shifts) may be derived and useful chemical shift ranges may be calculated.

No doubt more sophisticated approaches for both chemical calculation and chemical shift validation will eventually be developed (as they need to be), however, it is our hope that RefDB and its associated software will at least initiate a concerted movement towards improving the quality of data that NMR spectroscopists deposit in the field of biomolecular NMR.

3.5 Availability

RefDB, along with web-server versions of SHIFTX and SHIFTCOR are freely available at <http://redpoll.pharmacy.ualberta.ca>

Chapter 4

General Conclusions and Future Directions

The human genome project has revolutionized the life sciences. Through the development of a variety of high throughput technologies such as multiplexed DNA sequencing, soft-ionization mass spectrometry, DNA microarrays, and a host of other tools we have been able to fully sequence more than 60 organisms, including humans, mice, fish, fruit flies, nematodes and yeast (McPherson et al., 2001; Venter et al., 2001). These complex eukaryotic organisms are of considerable interest to pharmaceutical scientists as their genomic sequences provide exquisite insight into how they act and react. It is widely expected and, in fact, it has already been shown that detailed knowledge of the constituent sequences of pathogens and their hosts can accelerate the discovery of new drugs and new drug targets (Kinzler and Vogelstein, 1996). Yet, despite the abundance of DNA sequence data and the exciting promise of new drugs and drug targets to be discovered from this sequence data, there remain a number of significant hurdles still to be cleared.

For instance, even though we have sequenced nearly a dozen eukaryotic genomes, we still have considerable difficulty enumerating and locating all of their genes. This is quite problematic, especially for human and mouse genomes as it makes it exceedingly difficult to identify the genetic source of a disease or the appropriate target gene (or gene product) for drug development. An equally difficult problem for pharmaceutical

researchers (and probably for all life scientists in general) is concerned with how to track, update and correct the torrent of biological data that is pouring forth every day from genome centres, structural biology labs and proteomic nodes located around the globe. Manual data entry, validation and updating is simply too tedious and too time consuming for anyone to pursue. While these problems of gene identification and database development are but two among hundreds of "key" issues confronting life science researchers, I have chosen to address these two specific issues for this thesis.

In Chapter 2, I described the development (in collaboration with Dr. Peter Hooper) of a suite of computer programs for the identification of eukaryotic genes. Eukaryotic gene prediction is a particularly challenging task as the signals used by gene processing enzymes are poorly defined and highly dependent on context and patterns. The approach we developed, which is based on combining advanced statistical methods with sophisticated comparative sequence methods, appears to perform at a level greater than or equal to the best gene prediction programs so far published. Specifically, the GRPL and GRPL+ software packages make use of data training sets, reference point logistic classification (Hooper, 1999), generalized Hidden Markov Models (Michalski et al., 1998), dynamic programming (Needleman and Wunsch, 1970) and sequence database comparison (Altschul et al., 1997) to train, identify and refine the location of exons, introns and full-length genes. Our tests indicate that the software is able to predict the location of genes in organisms as diverse as humans, flies and plants with a correlation coefficient (at the nucleotide level) of 0.97. We also demonstrated that sequence database comparison (the + in GRPL+) was able to improve the performance of

all gene predictions, regardless of the quality of the initial prediction. Our tests also revealed the difficulty in correctly predicting gene locations as the size of the intergenic or intronic regions increased. This particular problem was not previously considered to be significant until the first draft of the human genome sequence was published in 2001 (Venter et al., 2001). This revealed the remarkable dispersion of exons and introns throughout the human genome and underlined the need for even more accurate gene prediction algorithms.

In an effort to improve GRPL and GRPL+ we may apply DNA sequence alignment methods against the expressed sequenced tag (EST) database. The EST database is a cDNA sequence database containing short, (300-600 base) sequence stretches at the 3' end of transcribed genes. Since these short sequences likely contain multiple exons, we expect to be able to improve GRPL's prediction of exon boundaries (at least for the 3' ends).

In addition to the use of other databases, it may also be possible to improve GRPL through the use of more rapid database search techniques. In particular, GRPL+ used a relatively slow alignment program called FASTALIGN (Wishart et al., 1994) to seek out sequence homologues. With recent improvements to BLAST (Altschul et al., 1997) we expect that the incorporation of this alignment algorithm should improve not only the search speed (by a factor of 4 or more) but also the sensitivity of the GRPL+ search. Similarly, the use of BLAST for the dbEST search should also enhance the speed and sensitivity.

In chapter 3 of this thesis, I described the development of an automated, self-updating, self-correcting database for biological (esp. chemical shift) data. Specifically I developed a series of software packages (SHIFTX, SHIFTCOR and UPDATE) along with a series of CGI scripts to perform web-based data mining, file comparison, data evaluation, data validation, file concatenation, file re-formatting and web-based file display. The results of this work appear in the database called RefDB – a database of reference-corrected chemical shifts specifically for peptides and proteins.

At the time of this writing, RefDB consists of 263 different proteins out of a total of ~400 fully (>80 % complete) assigned, non-redundant proteins in the BioMagResBank. Of these 263 proteins, 41 contain only ^1H assignments, 43 have ^{15}N and ^1H assignments and 178 proteins have ^1H , ^{13}C and ^{15}N assignments. Of those proteins with reported ^{13}C assignments, 97% have at least $^{13}\text{C}\alpha$ shift assignments, 85% have both $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ shift assignments while just 55% have ^{13}CO shift assignments. A total of 121 proteins have at least one corresponding X-ray structure while 142 have only NMR derived structures. During our analysis and corrections for RefDB we found that 60/212 (28.3%) of proteins with ^{15}N assignments required reference adjustments (up or down) of more than 1 ppm. Furthermore, 45/178 (25.3%), 51/161 (31.7%) and 23/105 (21.9%) of proteins in RefDB required reference adjustments of more than 0.5 ppm for their reported $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$ and ^{13}CO assignments, respectively. Overall, we found that the primary or "source" database for RefDB (the BioMagResBank) had more than 25% of its entries incorrectly or improperly referenced. We also found a number of obvious mis-

assignments or typographical errors that had been missed by BMRB depositors or BMRB data checkers. These discoveries are somewhat alarming as they indicate an unacceptably high level of non-compliance or shoddy experimental practices in biomolecular NMR. It is hoped that through the widespread adoption and use of tools such as SHIFTCOR and RefDB that this sorry state of affairs may soon be corrected.

While I chose to apply this self-updating database concept to the relatively narrow area of biological NMR, I believe the concepts and software tools developed for the RefDB project are highly generalizable. It is possible that these concepts and principles could be applied to a wide variety of biological or chemical data including sequence data, structural data, microarray data or any other biologically interesting data which is easily accessible from the web and for which a high percentage (>1%) of data entry errors is expected.

Appendix A

SimPRED

Introduction

For a number of years it has been observed that the three dimensional conformations of polypeptides, carbohydrates and nucleic acids had a weak to be observable effect on chemical shifts (Wishart et al., 1991; Wishart et al., 1992; Le and Oldfield, 1994; Osapay and Case, 1994; Wishart and Sykes, 1994a; Wishart and Sykes, 1994b; Oldfield, 1995). These effects are likely due to magnetic anisotropies arising from the asymmetric electron distribution found in many of these molecules. In 1992, Wishart *et al.* demonstrated that these so-called secondary chemical shifts were related to protein secondary structure.

Proteins are known to have two characteristic subsets of secondary structure – alpha helices and beta strands. These structures arise from the rotation of backbone *phi* and *psi* angles into regular or repeating patterns. These repeated patterns in *phi* and *psi* angles actually show up as identifiable patterns in chemical shift. Evidently, the magnetic anisotropy in the peptide bond is different enough between helices and beta strands to manifest itself as either an upfield shift (for helices) or a downfield shift (for beta-strands) in the α - ^1H resonances of amino acid residues (Wishart et al, 1992). These chemical shift changes have also been observed for ^{13}C and ^{15}N shifts of amino acids as well (Le and Oldfield, 1994; Wishart and Sykes, 1994b). By plotting the difference between the chemical shifts arising from the secondary structure and the chemical shifts expected in a

random coil, one can get a simple plot (called the chemical shift index), which shows the location and length of helices and beta-strands in a protein or polypeptide (Wishart et al., 1992; Wishart and Sykes, 1994b).

In a like manner, if one knows that sequence and secondary structure of a peptide or protein (from CD, from x-ray crystallography or from secondary structure prediction), then it stands to reason that one can predict the chemical shift of a given peptide or protein. The SimPRED program implements this hypothesis and allows researchers to essentially predict ^1H , ^{13}C and ^{15}N shifts using only sequence and secondary structure as input. In this way, SimPRED may assist researchers in the difficult process of spectral assignment.

Methods and Materials

SimPRED uses a set of residue-specific random coil ^1H , ^{13}C and ^{15}N chemical shift table (Wishart et al., 1995) along with nearest neighbor effects for ^{15}N and ^{13}CO resonances (Wishart et al, 1995; Wuthrich, 1994) and residue specific secondary shifts (Wishart et al., 1991). The predicted chemical shifts for each residue are calculated by adding the neighborhood sequence information to the corresponding secondary chemical shifts. SimPRED does not attempt to predict the ^1H .shifts of aromatic protons.

The SimPRED program is written by standard C language and has been tested on SGI, SunOS and Linux operating systems. SimPRED has also been implemented as a

Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://webmail.pharmacy.ualberta.ca/simpred/>

SimPred Version 1.0

This program predicts ^1H , ^{13}C and ^{15}N backbone chemical shifts of peptides and proteins using only the amino acid sequence and the predicted (or known) secondary structure as input. SimPred uses residue-specific random coil secondary shifts in its calculations.

To operate this server:

- 1) Decide which shifts you would like predicted.
- 2) Choose a local sequence/structure file using either the file browser or by typing/pasting the file in the text box.
- 3) Press submit.

For more information on running SimPred click this button: [HELP](#)

Predict:

Select desired file: [Browse...](#)

(Note: the input file must be in a specific format in order for this form to work. Refer to the [HELP](#) button above.)

OR type (paste) the sequence/structure file into the space below (see [HELP](#) for proper format):

[Submit](#) [Clear](#)

Problems? Questions? Suggestions? Please contact [Haiyan Zhang](#) or [David Wishart](#)

Done Internet

Figure A.1 The interface of SimPRED web server.

web server and is located at <http://redpoll.pharmacy.ualberta.ca/simpred>. Figure A.1 show a screen shot of the SimPRED server.

Input File Format

The user must provide a file containing both the protein sequence and its secondary structure. The file should only contain one protein sequence. The format of the file is similar to the FASTA format, the sequence given in standard IUPAC single letter code with a title marked by a '>'. Unlike the FASTA format, the secondary structure is marked on separate lines below the sequence. For secondary structure description, we use three letters: C -- Coil, B-- Beta Sheet, H -- Alpha Helix. An example of the input file is shown below:

```
>FRUCTOSE-1,6-BISPHOSPHATASE
ADQAPFDTDVVTLTRFVMEEGRKARGTGELTQLLNSLCTAVKAISSAVRK
CCCCCCCCBBBHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHC

AGIAHLYGIAGSTNVTGDQVKKLDVLSNDLVMNMLKSSFATCVLVSEEDK
CCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHCCBBB BBBCCC

HAIIVEPEKRGKYVVCFDPLDGSSNIDCLVSVGTIFGIYR
CBBBBCCCBBBBBBBBBBBCCCCCHHCCCCBBBBBBBBBBB
```

Note: There should be no space between the first letter and the last letter of each line. It is not necessary to have a blank line between each line of text.

Output Format

SIMPRED generates text format output file. The residue number is given in the left-most column, the amino acid sequence is given in the next column and the remaining chemical shifts are given in the other columns. The ¹H shifts proceed from HN to HA to HB to HG

to HD etc. depending on the residue type. The number 9999 indicates there is no chemical shift available for this type of nucleus. An example output file is shown as below:

No.	AA	CA	CB	CO	N	HN	HA	HB	HX
1	M	55.4	32.9	176.3	9999	9999	4.48	2.60	2.54	2.11 2.10 2.01
2	Q	54.3	31.5	174.5	122.3	8.51	4.88	2.25	2.10	2.07 1.88
3	I	60.0	40.0	175.0	124.6	8.58	4.61	1.76	1.26	1.10 0.77 0.61
4	F	56.3	41.7	173.5	125.6	8.78	5.08	3.01	2.71	
5	V	60.4	34.1	174.6	120.8	8.54	4.56	1.92	0.86	0.69
....										
....										

Results

Figure A.2 shows how SimPRED can be run on a Unix system.

Availability:

The SimPRED web server is on <http://redpoll.pharmacy.ualberta.ca/simpred>.

```

*****
* Package....:      SimPRED Version 1.0                      *
* Date.....:       May 27, 1999                             *
* Author.....:      Haiyan Zhang, David S. Wishart           *
* Purpose....:      To predict chemical shifts using         *
*                   sequence + secondary structures           *
*****

```

- 1) Predict all shifts
- 2) Predict ¹³C/¹⁵N shifts only
- 3) Predict ¹H shifts only
- 4) Help
- 5) Exit

```
>>1                                ← user input
```

Input Sequence Filename ("q" to quit)

```
>>test.seq                        ← user input
```

Enter output Filename

```
>>test.out                        ← user input
```

Thank you for using SimPRED!

Figure A.2 The procedure of running SimPRED on a Unix system. SimPRED is initiated by typing **simpred**. The places where the user needs to input information is indicated by 'user input' in bold face.

Appendix B

SHIFTOR

Introduction

^1H , ^{13}C and ^{15}N chemical shifts of amino acid residues in proteins have long recognized to be sensitive to local conformation (Wishart et al., 1991; Osapay and Case, 1991; Wishart and Sykes, 1994b). The correlation between secondary chemical shifts and backbone phi/psi angles is also well known (Spera and Bax, 1991; Wishart and Nip, 1998). As a result, several methods and a number of computer programs have been developed to predict protein secondary structures from chemical shifts (Wishart et al., 1992; Luginbuhl et al., 1995; Sharma and Rajarathnam, 2000). Given that we can accurately identify protein secondary structures from chemical shifts, it stands to reason that we should also be able to predict polypeptide phi/psi dihedral angles from chemical shifts as well. While this may seem like a trivial extension, it was only recently that this was attempted – using a program called TALOS.

TALOS (Cornilescu et al., 1999) uses sequence similarity to effect backbone dihedral prediction, with an accuracy of +/- 15°. Specifically, TALOS compares a query protein (and its associated chemical shifts) to a database of previously assigned proteins, including their sequence, their chemical shifts and their corresponding backbone dihedral angles (as determined by X-ray crystallography). TALOS uses a very simple measure of sequence similarity to predict the most likely backbone dihedral angles from homologous peptides (based on a combined measure of sequence similarity and chemical shift

similarity). In this way TALOS offers a simple, intuitive approach to converting raw chemical shift information into useful structural restraints for NMR-based structure generation and refinement. While TALOS has helped immeasurably in the application of chemical shifts to protein structure generation, it is a remarkably slow and inefficient program. Typically, to predict the phi/psi angles for a 100 residue protein requires several CPU hours. Furthermore, TALOS is incapable of recognizing the presence of homologous proteins in the databank and so its performance is less than satisfactory for predicting phi/psi angles for homologous (>35% ID) proteins. Here we present a far more efficient approach to backbone dihedral angle prediction called SHIFTOR. While still using the concept of database comparison as TALOS, SHIFTOR is able to predict the phi/psi angles of a 100 residue protein almost 200 times faster than TALOS.

Materials and Methods

Database construction

The SHIFTOR chemical shift database consists of 27 proteins comprising 3809 residues. All 27 proteins have fully assigned ^1H , ^{13}C and/or ^{15}N shifts and all have a corresponding high resolution (<2.2Å) X-ray structure. The phi and psi angles for each residue in the SHIFTOR database were calculated from the corresponding X-ray structure using VADAR (Wishart et al., 1995). All of the chemical shifts were further re-referenced and corrected using the SHIFTCOR program (Chapter 3).

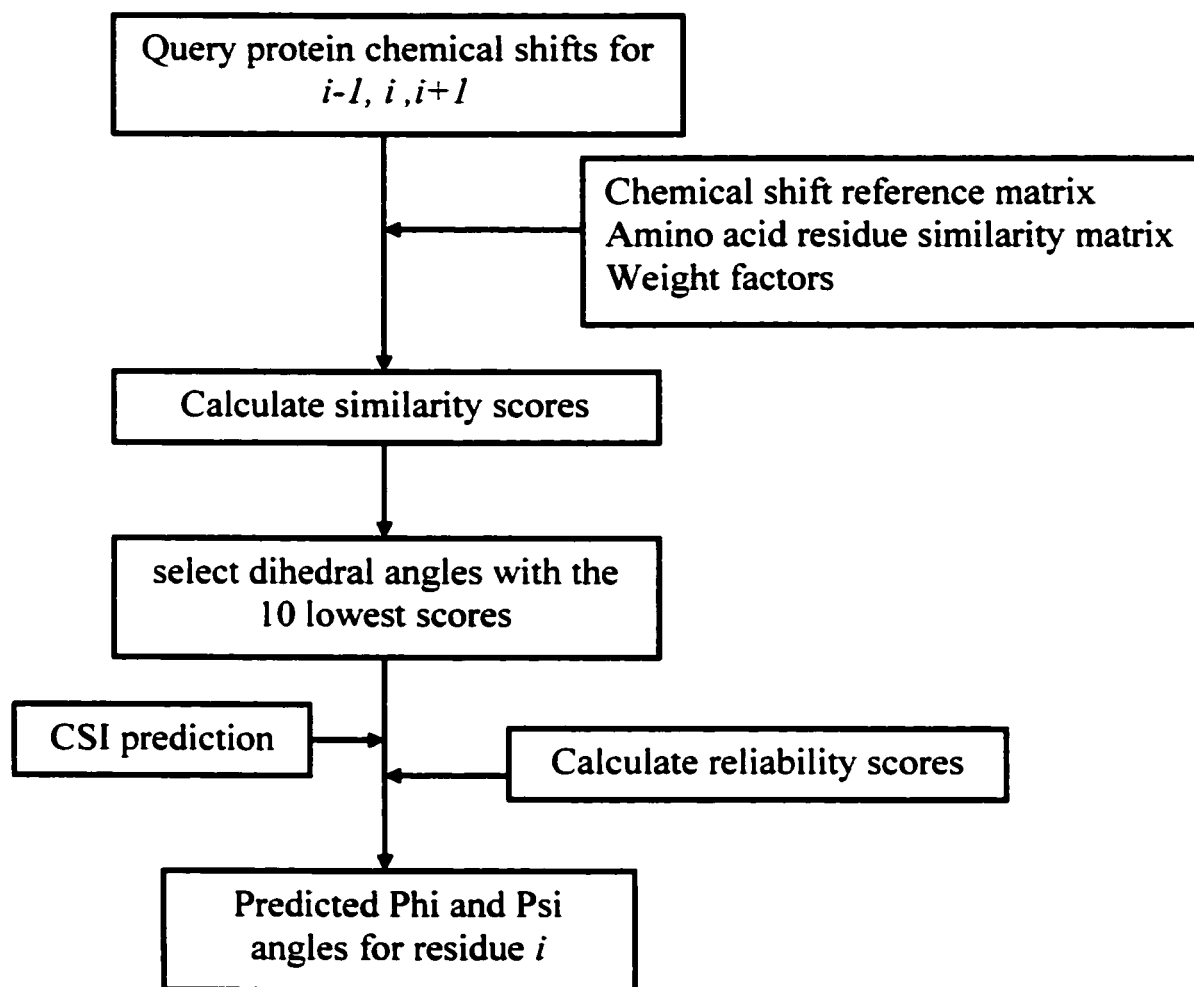


Figure B.1 The SHIFTOR flow chart.

Table B.1 Empirically optimized scaling factors (k_n^m) where m denotes homology ($^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}Co , ^{15}N , ^1HN) for weighting the relative importance of a given chemical shift or residue type in determining the SHIFTOR similarity score $S(i,j)$ (Equation B.1).

Residue	Homology	$^1\text{H}\alpha$	$^{13}\text{C}\alpha$	$^{13}\text{C}\beta$	^{13}Co	^{15}N	^1HN
n=-1	0.5	37	11	9	5	1	1
n=0	2.5	31	14	14	6	1.5	0.25
n=1	1.5	37	7	7	4	2	1.5

Similarity score calculations

SHIFTOR uses a sliding window technique to measure the similarity of a sequential set of three (a triplet) amino acids and their corresponding secondary chemical shifts for a given query protein against all triplets contained in the SHIFTOR database. For each query triplet with center residue i and each database triplet with center residue j , the similarity score $S(i,j)$ is calculated using Equation B.1. In this equation, K_n^m represents the empirically optimized scaling factors for weighting each type of chemical shift ($^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}Co , $^1\text{H}\alpha$, ^{15}N and ^1HN) and sequence similarity score. The values of K_n^m are displayed in Table B.1. For SHIFTOR we used the amino acid similarity scoring matrix developed by Wishart et al. (1994a) to calculate the similarity score (Σ_{ResType}) between any two corresponding residues. Note that Δ in Equation B.1 denotes the secondary chemical shifts, or the difference of the observed chemical shift from the random coil value (Wishart et al., 1995).

$$S(i,j) = \sum_{n=-1}^{+1} \left\{ 0.5 * K_n^0 \Sigma_{\text{ResType}} + K_n^1 (\Delta C\alpha_{i+n} - \Delta C\alpha_{j+n}) \right. \\
+ K_n^2 (\Delta C\beta_{i+n} - \Delta C\beta_{j+n}) \\
+ K_n^3 (\Delta C\alpha_{i+n} - \Delta C\alpha_{j+n}) \\
+ K_n^4 (\Delta H\alpha_{i+n} - \Delta H\alpha_{j+n}) \\
+ K_n^5 (\Delta N_{i+n} - \Delta N_{j+n}) \\
\left. + K_n^6 (\Delta HN_{i+n} - \Delta HN_{j+n}) \right\} \quad (B.1)$$

Description of the Database Search Procedure

An outline of the SHIFTOR search and calculation procedure is shown in Figure B.1. SHIFTOR begins by reading the sequence and assigned chemical shifts of the query protein. It then converts these raw shifts to secondary chemical shifts by subtracting the corresponding random coil values as given by Wishart and Nip (1998). A sliding window consisting of three sequential residues in the query protein (including secondary chemical shifts) is searched against every set of three residues (and their corresponding secondary shifts) in the SHIFTOR database. The combined chemical shift and residue similarity scores between the query triplet with database triplets is sequentially calculated for all 3807 triplets in the database and the ten triplets with the lowest scores are selected. The phi and psi dihedral angles for each central residue of the ten triplets is also extracted to help estimate of the most likely phi/psi angle of the query residue. In addition, a dihedral angle reliability score for each of the ten triplets is calculated using Equation B.2. Then the phi and psi angles are clustered, using a simple hierarchical clustering algorithm, by evaluating the difference between the 10 sets of predicted dihedral angles. Clusters are grouped if the difference of their phi or psi angles is less than 15°. The reliability score for each group is the sum of the individual reliability score for each triplet. The group

#NUM	AA	HA	CA	CB	CO	N	HN
1	M	4.22	54.2	32.9	170.9	0	0
2	Q	5.27	54.9	30.4	176.4	124.5	8.96
3	I	4.21	59.4	41.7	172.7	116.7	8.35
4	F	5.61	54.8	40.9	175.5	120.1	8.63
5	V	4.79	60.2	33.8	175.2	122.8	9.33
6	K	5.34	54.3	34.1	177.5	129.4	8.98
7	T	5.00	60.2	70.3	177.3	116.9	8.77
8	L	4.33	57.3	41.7	179.2	122.8	9.13
9	T	4.41	61.2	68.8	175.9	107.4	7.67
10	G	4.36	45.0	0	174.4	110.7	7.86
11	K	4.37	56.1	33.2	176.1	123.4	7.30
12	T	5.08	62.1	69.6	174.7	122.1	8.65
13	I	4.53	59.8	40.5	175.6	129.1	9.57
14	T	4.97	61.8	69.5	174.1	123.2	8.75
15	L	4.75	52.5	46.6	174.9	126.7	8.77
...							

Figure B.2 An example of the SHIFTOR program input formats. The first line of the input file must begin with '#' and must provide descriptions for each column (NUM - number of residue; AA - amino acid residue name; HA, CA, CB, CO, N and HN - six kinds of backbone chemical shifts). The columns are separated by at least one space. The number of columns may vary from 3 - 8. The minimum requirement is one 'NUM' column, one 'AA' column and one of chemical shift column. The order of the columns may vary. The amino acid entry should in format of IUPAC single letter code. For unassigned chemical shifts, a '0' value should be entered.

R(phi)	R(psi)	Res	PHI	PSI
1.00	0.70	Q	-91.00	138.30
0.60	0.90	I	-131.10	163.00
0.90	0.90	F	-116.00	140.20
0.90	1.00	V	-118.00	114.20
0.70	0.80	K	-95.20	127.50
0.70	0.40	T	-99.60	170.80
0.90	0.70	L	-73.40	-6.90
0.90	0.80	T	-101.40	14.90
0.70	0.60	G	77.40	16.50
0.70	0.80	K	-96.30	138.10
0.60	0.70	T	-119.90	131.80
0.70	0.50	I	-109.50	142.00
0.70	1.00	T	-101.40	139.70
0.60	0.70	L	-126.40	154.00
0.70	0.90	F	-111.80	121.10
0.70	0.80	V	-139.00	170.70
		...		

Figure B.3 An example of the SHIFTOR output format. All output files have five columns. The middle column is the amino acid residue (one letter code). The two columns on the right are the predicted phi and psi angles, respectively. The two columns on the left are the reliability scores for the phi and psi angle predictions. The reliability score varies between 0~1 with 1 being most reliable and 0 being least.

with highest reliability score is then selected and the mean phi and psi angle for that cluster is used as the predicted value for the central residue of the query triplet. To prevent spurious predictions, chemical shift indices (Wishart et al., 1992; Wishart and Sykes, 1994c) are also calculated from the query protein shifts to help choose the correct torsion angles.

$$R = 2^{-(3*S(i,j)/238.5)^2} \quad (B.2)$$

Results

Examples of the input and output file formats for a typical SHIFTOR prediction are shown in Figure B.2 and Figure B.3 respectively. As can be seen from these figures SHIFTOR provides predicted values for phi and psi angles, as well as a reliability score for the predicted results (Figure B.3).

Our results indicate that SHIFTOR is able to correctly predict (+/- 15°) more than 80% of a given query protein's phi and psi angles. Because we have tried to include global sequence bias into the algorithm we have found that SHIFTOR is able to correctly

Table B.2 Performance comparison between SHIFTOR and TALOS on 4 proteins (400 residues). The phi and psi prediction for a residue is considered to be corrected if the difference of their values to the actual values is less than 30.

Programs	Results without query sequence in the database (% correct)	Results with query sequence in the database (% correct)
SHIFTOR	82%	95%
TALOS	77.8%	81%

predict ~95% of a query protein's torsion angles if the query protein (or a close homologue) is already in the database. This high performance should be expected as any reasonable dihedral angle prediction algorithm should be able to “recover” the dihedral angles it was based on. As seen in Table B.2, the accuracy of the results we obtain with SHIFTOR are at least comparable and generally superior to those obtained with TALOS. However, the most noteworthy improvement by SHIFTOR over TALOS is in its calculation speed (20 seconds vs. 2 hours). This speed-up was possible due to more efficient use of memory, the use of improved (i.e. more efficient) alignment algorithms and the use of a compiled language (C – for SHIFTOR) instead of an interpretive language (Tcl/Tk for TALOS).

Availability:

The SHIFTOR web server is located at <http://redpoll.pharmacy.ualberta.ca/shifor>.

Reference

- Abagyan, R.A. and Batalov, S. (1997). "Do aligned sequences share the same fold?" *J Mol Biol* 273(1): 355-68.
- Akutsu, T., Miyano, S. and Kuhara, S. (2000). "Inferring qualitative relations in genetic networks and metabolic pathways." *Bioinformatics* 16(8): 727-34.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). "Basic local alignment search tool." *J Mol Biol* 215(3): 403-10.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* 25(17): 3389-402.
- Appel, R.D., Bairoch, A. and Hochstrasser, D.F. (1994). "A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server." *Trends Biochem Sci* 19(6): 258-60.
- Attwood, T.K., Beck, M.E., Flower, D.R., Scordis, P. and Selley, J.N. (1998). "The PRINTS protein fingerprint database in its fifth year." *Nucleic Acids Res* 26(1): 304-8.
- Bairoch, A. and Apweiler, R. (2000). "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000." *Nucleic Acids Res* 28(1): 45-8.
- Bairoch, A., Bucher, P. and Hofmann, K. (1997). "The PROSITE database, its status in 1997." *Nucleic Acids Res* 25(1): 217-21.
- Barker, W.C., Garavelli, J.S., Hou, Z., Huang, H., Ledley, R.S., McGarvey, P.B., Mewes, H.W., Orcutt, B.C., Pfeiffer, F., Tsugita, A., Vinayaka, C.R., Xiao, C., Yeh, L.S. and Wu, C. (2001). "Protein Information Resource: a community resource for expert annotation of protein data." *Nucleic Acids Res* 29(1): 29-32.
- Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000). "The Pfam protein families database." *Nucleic Acids Res* 28(1): 263-6.
- Bax, A. and Subramanian, A.B. (1986). "Sensitivity-enhanced two-dimensional heteronuclear shift correlation NMR spectroscopy." *J. Magn. Reson.* 67: 565-569.
- Beger, R.D. and Bolton, P.H. (1997). "Protein phi and psi dihedral restraints determined from multidimensional hypersurface correlations of backbone chemical shifts and their use in the determination of protein tertiary structures." *J Biomol NMR* 10(2): 129-42.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L.

(2000). "GenBank." *Nucleic Acids Res* 28(1): 15-8.

Berman, H.M., Olson, W.K., Beveridge, D.L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.H., Srinivasan, A.R. and Schneider, B. (1992). "The nucleic acid database. A comprehensive relational database of three- dimensional structures of nucleic acids." *Biophys J* 63(3): 751-9.

Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gillicand, G., Weissig, H., and Westbrook, J. (2000) . ""The Protein DataBank and the challenge of structural genomics." *Nat Struct Biol* 7 Suppl: 957-9

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures." *J Mol Biol* 112(3): 535-42.

Bjorndahl, T.C., Watson, M.S., Slupsky, C.M., Spyropoulos, L., Sykes, B.D. And Wishart, D.S. (2001). "Complete ¹H, ¹³C and ¹⁵N backbone assignments for the hepatitis A virus 3C protease." *J Biomol NMR* 19(2): 187-8

Bleasby, A.J., Akrigg, D. and Attwood, T.K. (1994). "OWL--a non-redundant composite protein sequence database." *Nucleic Acids Res* 22(17): 3574-7.

Bork, P. and Gibson, T.J. (1996). "Applying motif and profile searches." *Methods Enzymol* 266: 162-84.

Borodovsky, M. and McIninch, J. (1993). "Recognition of genes in DNA sequence with ambiguities." *Biosystems* 30(1-3): 161-71.

Borodovsky, M.Y. and McIninch, J.D. (1993). "GENMARK: Parallel gene recognition for both DNA strands." *Comput. & Chem.* 17: 123-133.

Brenner, S.E., Chothia, C. and Hubbard, T.J. (1998). "Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships." *Proc Natl Acad Sci U S A* 95(11): 6073-8.

Burge, C. and Karlin, S. (1997). "Prediction of complete gene structures in human genomic DNA." *J Mol Biol* 268(1): 78-94.

Burset, M. and Guigo, R. (1996). "Evaluation of gene structure prediction programs." *Genomics* 34(3): 353-67.

Buttery, J.P. and Moxon, E.R. (2000). "Designing meningitis vaccines." *J R Coll Physicians Lond* 34(2): 163-8.

Case, D.A. (2000). "Interpretation of chemical shifts and coupling constants in macromolecules." *Curr Opin Struct Biol* 10(2): 197-203

- Claverie, J.M. and States, D. (1993). "Information enhancement methods for large scale sequence analysis." *Comput. & Chem.* 20(119-122).
- Clore, G.M. and Gronenborn, A.M. (1998). "New methods of structure refinement for macromolecular structure determination by NMR." *Proc Natl Acad Sci U S A* 95(11): 5891-8.
- Cohen, A.S., Najarian, D.R. and Karger, B.L. (1990). "Separation and analysis of DNA sequence reaction products by capillary gel electrophoresis." *J Chromatogr* 516(1): 49-60.
- Cornilescu, G., Delaglio, F. and Bax, A. (1999). "Protein backbone angle restraints from searching a database for chemical shift and sequence homology." *J Biomol NMR* 13(3): 289-302.
- Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data* 2nd Edition. London: Chapman & Hall.
- Dayhoff, M.O., Eck, R.V., Chang, M.A., Sochard, M.R. (1965). "Atlas of Protein Sequence and Structure." *National Biomedical Research Foundation, Silver Spring MD*.
- Deloukas, P., Schuler, G.D., Gyapay, G., Beasley, E.M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matisse, T.C., McKusick, K.B., Beckmann, J.S., Bentolila, S., Bihoreau, M., Birren, B.B., Browne, J., Butler, A., Castle, A.B., Chiannilkulchai, N., Clee, C., Day, P.J., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Bentley, D.R. and et al. (1998). "A physical map of 30,000 human genes." *Science* 282(5389): 744-6.
- DesJarlais, R.L., Seibel, G.L., Kuntz, I.D., Furth, P.S., Alvarez, J.C., Ortiz de Montellano, P.R., DeCamp, D.L., Babe, L.M. and Craik, C.S. (1990). "Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease." *Proc Natl Acad Sci U S A* 87(17): 6644-8.
- Discala, C., Benigni, X., Barillot, E. and Vaysseix, G. (2000). "DBcat: a catalog of 500 biological databases." *Nucleic Acids Res* 28(1): 8-9.
- Doolittle, R.F. (1997). "Some reflections on the early days of sequence searching." *J Mol Med* 75(4): 239-41.
- Drews, J. (2000). "Drug discovery: a historical perspective." *Science* 287(5460): 1960-4.
- Drews, J. and Ryser, S. (1997). "The role of innovation in drug development." *Nat Biotechnol* 15(13): 1318-9.
- Fesik, S.W. (1993). "NMR structure-based drug design." *J Biomol NMR* 3(3): 261-9.
- Fickett, J.W. (1982). "Recognition of protein coding regions in DNA sequences." *Nucleic Acids Res* 10(17): 5303-18.

- Friedrich, G.A. (1996). "Moving beyond the genome projects." *Nat Biotechnol* 14(10): 1234-7.
- Wagner, G., Pardi, A. and Wuthrich, K. (1983). *J. Am. Chem. Soc* 105: 5948.
- Gardner, K.H., Rosen, M.K. and Kay, L.E. (1997). "Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR." *Biochemistry* 36(6): 1389-401.
- Gelfand, M.S. (1995). "Prediction of function in DNA sequence analysis." *J Comput Biol* 2(1): 87-115.
- Gish, W. and States, D.J. (1993). "Identification of protein coding regions by database similarity search." *Nat Genet* 3(3): 266-72.
- Gracy, J. and Argos, P. (1998). "DOMO: a new database of aligned protein domains." *Trends Biochem Sci* 23(12): 495-7.
- Gribskov, M., Devereux, J. and Burgess, R.R. (1984). "The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression." *Nucleic Acids Res* 12(1 Pt 2): 539-49.
- Gronenborn, A.M. and Clore, G.M. (1994). "Identification of N-terminal helix capping boxes by means of ¹³C chemical shifts." *J Biomol NMR* 4(3): 455-8.
- Gronwald, W., Willard, L., Jellard, T., Boyko, R.F., Rajarathnam, K., Wishart, D.S., Sonnichsen, F.D. and Sykes, B.D. (1998). "CAMRA: chemical shift based computer aided protein NMR assignments." *Journal of Biomolecular NMR* 12(3): 395-405.
- Guigo, R., Knudsen, S., Drake, N. and Smith, T. (1992). "Prediction of gene structure." *J Mol Biol* 226(1): 141-57.
- Haigh, C.W. and Mallion, R.B. (1979). "Ring Current Theories in NMR." *Prog. Nucl. Magn. Reson. Spectrosc.* 13: 303.
- Henikoff, S., Henikoff, J.G. and Pietrokovski, S. (1999). "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations." *Bioinformatics* 15(6): 471-9.
- Hooper, P.M. (1999). "Reference point logistic classification." *J. Class.* 16: 91-116.
- Hooper, P.M. (2001). "Reference point logistic regression and the identification of DNA functional sites." *J. Class.* 18:81-107.
- Hooper, P.M., Zhang, H. and Wishart, D.S. (2000). "Prediction of genetic structure in eukaryotic DNA using reference point logistic regression and sequence alignment."

Bioinformatics 16(5): 425-38.

Hutchinson, G.B. and Hayden, M.R. (1992). "The prediction of exons through an analysis of spliceable open reading frames." *Nucleic Acids Res* 20(13): 3453-62.

Iwadata, M., Asakura, T. and Williamson, M.P. (1999). "C alpha and C beta carbon-13 chemical shifts in proteins from an empirical database." *J Biomol NMR* 13(3): 199-211.

Jenuth, J.P. (2000). "The NCBI. Publicly available tools and resources on the Web." *Methods Mol Biol* 132: 301-12.

Kinzler, K.W. and Vogelstein, B. (1996). "Lessons from hereditary colorectal cancer." *Cell* 87:159-170.

Kneale, G.G. and Kennard, O. (1984). "The EMBL nucleotide sequence data library." *Biochem Soc Trans* 12(6): 1011-4.

Krogh, A. (1997). "Two methods for improving performance of an HMM and their application for gene finding." *Proc Int Conf Intell Syst Mol Biol* 5: 179-86.

Kulp, D., Haussier, D., Reese, M.G. & Eeckman, F.H. (1996). "A generalized Hidden Markov Model for the recognition of human genes in DNA." *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*. AAI Press, Menlo Park, CA.

Kumar, R., Lumsden, A., Ciclitira, P.J., Ellis, H.J. and Laurie, G.W. (2000). "Human genome search in celiac disease using gliadin cDNA as probe." *J Mol Biol* 300(5): 1155-67.

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1995). "The impact of direct refinement against proton chemical shifts on protein structure determination by NMR." *J Magn Reson B* 107(3): 293-7.

Kuszewski, J., Qin, J., Gronenborn, A.M. and Clore, G.M. (1995). "The impact of direct refinement against ¹³C alpha and ¹³C beta chemical shifts on protein structure determination by NMR." *J Magn Reson B* 106(1): 92-6.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993). "PROCHECK: a program to check the stereochemical quality of protein structures." *J. Appl. Cryst.* 26:283-291.

Le, H. and Oldfield, E. (1994). "Correlation between ¹⁵N NMR chemical shifts in proteins and secondary structure." *J Biomol NMR* 4(3): 341-8.

Levinthal, C. (1966). "Molecular model-building by computer." *Sci Am* 214(6): 42-52.

Levinthal, G. and Ray, M. (1996). "Hepatitis A: from epidemic jaundice to a vaccine-preventable disease." *Gastroenterologist* 4(2): 107-17.

Lipman, D.J. and Pearson, W.R. (1985). "Rapid and sensitive protein similarity searches." *Science* 227(4693): 1435-41.

Lopez, R., Larsen, F. and Prydz, H. (1994). "Evaluation of the exon predictions of the GRAIL software." *Genomics* 24(1): 133-6.

Luckey, J.A., Drossman, H., Kostichka, A.J., Mead, D.A., D'Cunha, J., Norris, T.B. and Smith, L.M. (1990). "High speed DNA sequencing by capillary electrophoresis." *Nucleic Acids Res* 18(15): 4417-21.

Luginbuhl, S., Szyperski, T. and Wuthrich, K. (1995). "Statistical basis for the use of $^{13}\text{C}\alpha$ chemical shifts in protein structure determination." *J. Magn. Reson.* 109:229-233

Manber, U. and Wu, S. (1994). "GLIMPSE: a tool to search through entire file system." *Usenix Winter 1994 Technical Conference* (best paper award), San Francisco (January 1994), pp. 23-32.

Manber, U. and Bigot, P. (1997). "The Search Broker." *USENIX Symposium on Internet Technologies and Systems (NSITS'97)*, Monterey, California, pp.231-239.

Marchington, T. (1995). "From data to drugs." *Biotechnology (N Y)* 13(3): 239-42.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wuthrich, K. (1998). "Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy." *J Biomol NMR* 12(1): 1-23.

Maurer, T. and Kalbitzer, H.R. (1996). "Indirect Referencing of ^{31}P and ^{19}F NMR Spectra." *J Magn Reson B* 113(2): 177-8.

Maxam, A.M. and Gilbert, W. (1977). "A new method for sequencing DNA." *Proc Natl Acad Sci U S A* 74(2): 560-4.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., Fulton, R., Kucaba, T.A., Wagner-McPherson, C., Barbazuk, W.B., Gregory, S.G., Humphray, S.J., French, L., Evans, R.S., Bethel, G., Whittaker, A., Holden, J.L., McCann, O.T., Dunham, A., Soderlund, C., Scott, C.E., Bentley, D.R., Schuler, G., Chen, H.C., Jang, W., Green, E.D., Idol, J.R., Maduro, V.V., Montgomery, K.T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J.H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P.J., Catanese, J.J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V.G., Kirsch, I.R., Reid, T.,

Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J.F., Hawkins, T., Myers, R.M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N.E., Cox, D.R., Haussler, D., Kent, W.J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X.N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H.S., Sakaki, Y., Shimizu, N., Asakawa, S., Kawasaki, K., Sasaki, T., Shintani, A., Shimizu, A., Shibuya, K., Kudoh, J., Minoshima, S., Ramser, J., Seranski, P., Hoff, C., Poustka, A., Reinhardt, R. and Lehrach, H. (2001). "A physical map of the human genome." *Nature* 409(6822): 934-41.

Mendelsohn, A.R. and Brent, R. (1999). "Protein interaction methods--toward an endgame." *Science* 284(5422): 1948-50.

Metzler, W.J., Constantine, K.L., Friedrichs, M.S., Bell, A.J., Ernst, E.G., Lavoie, T.B. and Mueller, L. (1993). "Characterization of the three-dimensional solution structure of human profilin: ¹H, ¹³C, and ¹⁵N NMR assignments and global folding pattern." *Biochemistry* 32(50): 13818-29.

Michalski, R., Bratko, I., and Bratko, A. (1998). "Machine Learning and data mining: methods and applications." Chichester, West Sussex, England.

Michie, D., Spiegelhalter, D.J., and Taylor, C.C. (1994). "Machine Learning, Neural and Statistical Classification." *New York: Ellis Horwood*.

Moseley, H.N. and Montelione, G.T. (1999). "Automated analysis of NMR assignments and structures from proteins." *Curr Opin Struct Biol* 9(5): 635-42

Moukheiber, Z. (1998). "Back to Nature." *Forbes Magazine*: 146-147.

Neal, S. and Wishart, D.S. (unpublished data).

Needleman, S.B. and Wunsch, C.D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *J Mol Biol* 48(3): 443-53.

Normile, D. (1999). "Building working cells 'in silico'." *Science* 284(5411): 80-1.

Oldfield, E. (1995). "Chemical shifts and three-dimensional protein structures." *J Biomol NMR* 5(3): 217-25.

Olson, A.J. and Goodsell, D.S. (1998). "Automated docking and the search for HIV protease inhibitors." *SAR QSAR Environ Res* 8(3-4): 273-85.

Osapay, K. and Case, D.A. (1994). "Analysis of proton chemical shifts in regular secondary structure of proteins." *J Biomol NMR* 4(2): 215-30.

Pearson, W.R. and Lipman, D.J. (1988). "Improved tools for biological sequence

- comparison." *Proc Natl Acad Sci U S A* 85(8): 2444-8.
- Pennisi, E. (1999). "Keeping genome databases clean and up to date." *Science* 286(5439): 447-50.
- Persidis, A. (1998). "The business of pharmacogenomics." *Nat Biotechnol* 16(2): 209-10
- Persidis, A. (1999). "Bioinformatics." *Nat Biotechnol* 17(18): 828-30
- Pons, J.L. and Delsuc, M.A. (1999). "RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins." *J Biomol NMR* 15(1): 15-26.
- Potts, B.C. and Chazin, W.J. (1998). "Chemical shift homology in proteins." *J. Biomol NMR* 11(1): 45-57
- Redfield, C. and Roberson, J. (1991). "Proceeding of a NATO advanced research workshop on computational aspects of the study of biological macromolecules by NMR." Plenum Press, New York, NY.
- Retief, J.D. (2000). "Phylogenetic analysis using PHYLIP." *Methods Mol Biol* 132: 243-58.
- Rogozin, I.B. and Milanesi, L. (1997). "Analysis of donor splice sites in different eukaryotic organisms." *J Mol Evol* 45(1): 50-9.
- Rost, B., Sander, C. and Schneider, R. (1994). "PHD--an automatic mail server for protein secondary structure prediction." *Comput Appl Biosci* 10(1): 53-60.
- Salzberg, S., Delcher, A.L., Fasman, K.H. and Henderson, J. (1998). "A decision tree system for finding genes in DNA." *J Comput Biol* 5(4): 667-80.
- Salzberg, S.L. (1997). "A method for identifying splice sites and translational start sites in eukaryotic mRNA." *Comput Appl Biosci* 13(4): 365-76.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1977). "Nucliotide sequence of bacteriophage phi X174 DNA." *Nature* 265(5596): 687-95.
- Sankoff, D. and Kruskal, J.B. (1983). "Time Wraps, String Edits and Macromolecules. The Theory and Practice of Sequence Comparison." *Reading, Massachussetts: Addison-Wesley*: 382 pages.
- Sanseau, P. (2001). "Impact of human genome sequencing for in silico target discovery." *Drug Discov Today* 6(6): 316-323.
- Schwarzinger, S., Kroon, G.J., Foss, T.R., Chung, J., Wright, P.E. and Dyson, H.J. (2001).

"Sequence-dependent correction of random coil NMR chemical shifts." *J Am Chem Soc* 123(13): 2970-8

Schwarzinger, S., Kroon, G.J., Foss, T.R., Wright, P.E. and Dyson, H.J. (2000). "Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView." *J Biomol NMR* 18(1): 43-8.

Searls, D.B. (2000). "Using bioinformatics in gene and drug discovery." *Drug Discov Today* 5(4): 135-143.

Searls, D.B. and Murphy, K.P. (1995). "Automata-Theoretic Models of Mutation and Alignment." *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*. AAAI Cambridge: pp. 341-349.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991). "A relational database for sequence-specific protein NMR data." *J Biomol NMR* 1(3): 217-36.

Sharma, D., Rajarathnam, K. (2000) "13C NMR chemical shifts can predict disulfide bond formation." *J Biomol. NMR*. 2000 Oct;18 (2):165-71.

Shuker, S.B., Hajduk, P.J., Meadows, R.P. and Fesik, S.W. (1996). "Discovering high-affinity ligands for proteins: SAR by NMR." *Science* 274(5292): 1531-4.

Sintchak, M.D. and Nimmesgern, E. (2000). "The structure of inosine 5'-monophosphate dehydrogenase and the design of novel inhibitors." *Immunopharmacology* 47(2-3): 163-84.

Smith, T.F. and Waterman, M.S. (1981). "Identification of common molecular subsequences." *J Mol Biol* 147(1): 195-7.

Snyder, E.E. and Stormo, G.D. (1995). "Identification of protein coding regions in genomic DNA." *J Mol Biol* 248(1): 1-18.

Solovyev, V. and Salamov, A. (1997). "The Gene-Finder computer tools for analysis of human and model organisms genome sequences." *Proc Int Conf Intell Syst Mol Biol* 5: 294-302.

Solovyev, V., Salamov, A. and Lawrence, C.B. (1994). "Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames." *Nucleic Acids Res* 22(24): 5156-63.

Spera, S. and Bax, A. (1991). "Empirical correlation between protein backbone conformation and C α and C β 13C nuclear magnetic resonance chemical shifts." *J. Am. Chem. Soc* 113: 5490-5492.

Staden, R. and McLachlan, A.D. (1982). "Codon preference and its use in identifying protein coding regions in long DNA sequences." *Nucleic Acids Res* 10(1): 141-56.

Sternlicht, H. and Wilson, D. (1967). "Magnetic resonance studies of macromolecules. I. Aromatic-methyl interactions and helical structure effects in lysozyme." *Biochemistry* 6(9): 2881-92.

Stevens, R., Goble, C., Baker, P. and Brass, A. (2001). "A classification of tasks in bioinformatics." *Bioinformatics* 17(2): 180-8.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Lombard, V., Lopez, R., Parkinson, H., Redaschi, N., Sterk, P., Stoehr, P. and Tuli, M.A. (2001). "The EMBL nucleotide sequence database." *Nucleic Acids Res* 29(1): 17-21.

Swerdlow, H. and Gesteland, R. (1990). "Capillary gel electrophoresis for rapid, high resolution DNA sequencing." *Nucleic Acids Res* 18(6): 1415-9.

Thomas, A. and Skolnick, M.H. (1994). "A probabilistic model for detecting coding regions in DNA sequences." *IMA J Math Appl Med Biol* 11(3): 149-60.

Uberbacher, E.C. and Mural, R.J. (1991). "Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach." *Proc Natl Acad Sci U S A* 88(24): 11261-5.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love,

A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001). "The sequence of the human genome." *Science* 291(5507): 1304-51.

Verweij, J. (1999). "Rational design of new tumoractivated cytotoxic agents." *Oncology* 57 Suppl 1: 9-15.

Vriend, G. (1990). "WHAT IF: A molecular modeling and drug design program." *J. Mol. Graph.* 8:52-56.

Wang, Y., Bjorndahl, T.C. and Wishart, D.S. (2000). "Complete ¹H and non-carbonylic ¹³C assignments of native hen egg-white lysozyme." *J Biomol NMR* 17(1): 83-4.

Wang, Y., Bjorndahl, T.C. and Wishart, D.S. (2000). "Letter to the editor: Complete ¹H and non-carbonylic ¹³C assignments of native hen egg-white lysozyme [letter] [In Process Citation]." *J Biomol NMR* 17(1): 83-4.

Wang, Y., Henz, M.E., Gallagher, N.L., Chai, S., Gibbs, A.C., Yan, L.Z., Stiles, M.E., Wishart, D.S. and Vederas, J.C. (1999). "Solution structure of carnobacteriocin B2 and implications for structure-activity relationships among type IIa bacteriocins from lactic acid bacteria." *Biochemistry* 38(47): 15438-47.

Wang, Y., Nip, A.M. and Wishart, D.S. (1997). "A simple method to quantitatively measure polypeptide JHNH alpha coupling constants from TOCSY or NOESY spectra." *J Biomol NMR* 10(4): 373-82.

Waterman, M.S. (1984). "Efficient sequence alignment algorithms." *J Theor Biol* 108(3): 333-7.

Wider, G. and Wuthrich, K. (1999). "NMR spectroscopy of large molecules and multimolecular assemblies in solution." *Curr Opin Struct Biol* 9(5): 594-601.

- Williamson, M.P. and Asakura, T. (1992). "The application of ^1H NMR chemical shift calculations to diastereotopic groups in proteins." *FEBS Lett* 302(2): 185-8.
- Williamson, M.P., Asakura, T., Nakamura, E. and Demura, M. (1992). "A method for the calculation of protein α -CH chemical shifts." *J Biomol NMR* 2(1): 83-98.
- Williamson, M.P., Kikuchi, J. and Asakura, T. (1995). "Application of ^1H NMR chemical shifts to measure the quality of protein structures." *J Mol Biol* 247(4): 541-6.
- Williamson, M.P. and Asakura, T. (1993). "Empirical comparisons of models for chemical shift calculation in proteins." *J. Magn. Reson. ser. B* 101: 63-71.
- Wilson, R.K., Chen, C. and Hood, L. (1990). "Optimization of asymmetric polymerase chain reaction for rapid fluorescent DNA sequencing." *Biotechniques* 8(2): 184-9.
- Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995). " ^1H , ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects." *J Biomol NMR* 5(1): 67-81.
- Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995). " ^1H , ^{13}C and ^{15}N chemical shift referencing in biomolecular NMR." *J Biomol NMR* 6(2): 135-40.
- Wishart, D.S., Boyko, R.F., Willard, L., Richards, F.M. and Sykes, B.D. (1994a). "SEQSEE: a comprehensive program suite for protein sequence analysis." *Computer Applications in the Biosciences* 10(2): 121-32.
- Wishart, D.S. and Case, D.A. (2001). "Use of chemical shifts in macromolecular structure determination." *Methods Enzymol.* 338:3-34
- Wishart, D.S., Fortin, S., Woloschuk, D.R., Wong, W., Rosborough, T., Van Domselaar, G., Schaeffer, J. and Szafron, D. (1997). "A platform-independent graphical user interface for SEQSEE and XALIGN." *Comput Appl Biosci* 13(5): 561-2.
- Wishart, D.S. and Nip, A.M. (1998). "Protein chemical shift analysis: a practical guide. [Review] [57 refs]." *Biochemistry & Cell Biology* 76(2-3): 153-63.
- Wishart, D.S. and Sykes, B.D. (1994b). "The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data." *J Biomol NMR* 4(2): 171-80.
- Wishart, D.S. and Sykes, B.D. (1994c). "Chemical shifts as a tool for structure determination." *Methods Enzymol* 239: 363-92.
- Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991). "Relationship between nuclear

magnetic resonance chemical shift and protein secondary structure." *J Mol Biol* 222(2): 311-33.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992). "The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy." *Biochemistry* 31(6): 1647-51.

Wishart, D.S., Watson, M.S., Boyko, R.F. and Sykes, B.D. (1997). "Automated ¹H and ¹³C chemical shift prediction using the BioMagResBank." *J Biomol NMR* 10(4): 329-36.

Wishart D.S., Willard L., Richards F.M. and Sykes B.D. (1994d) "VADAR: A comprehensive program for protein structure evaluation. Version 1.2." Edmonton, Alberta, Canada.

Wootton, J.C. and Federhen, S. (1993). "Statistics of local complexity in amino acid sequences and sequence database." *Comput. & Chem.* 17: 191-201.

Wuthrich, K. (1986). "NMR of proteins and Nucleic Acids." John Wiley & Sons, New York.

Xu, X. and Case, D.A. (2001). "Automated prediction of ¹⁵N, ¹³Ca, ¹³C β and ¹³C' chemical shifts in proteins using a density functional database." *Journal of Biomolecular NMR* 21(4): 321-333,

Xu, Y., Einstein, J.R., Mural, R.J., Shah, M. & Uberbacher, E.C. (1994). "An improved system for exon recognition and gene modeling in human DNA sequences." *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI press, Menlo Park, CA,; pp, 376-384.

Zhang, J. and Madden, T.L. (1997). "PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation." *Genome Res* 7(6): 649-56.