Bioinformatics Methods for NMR Chemical Shift Data

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik

der Ludwig-Maximilians-Universität München

Simon W. Ginzinger

vorgelegt am 24.10.2007

Erster Gutachter:
Prof. Dr. Volker Heun, Ludwig-Maximilians-Universität München
Zweiter Gutachter:
Prof. Dr. Robert Konrat, Universität Wien

Rigorosum: 8. Februar 2008

Kurzfassung

Die nukleare Magnetresonanz-Spektroskopie (NMR) ist eine der wichtigsten Methoden, um die drei-dimensionale Struktur von Biomolekülen zu bestimmen. Trotz großer Fortschritte in der Methodik der NMR ist die Auflösung einer Proteinstruktur immer noch eine komplizierte und zeitraubende Aufgabe. Das Ziel dieser Doktorarbeit ist es, Bioinformatik-Methoden zu entwickeln, die den Prozess der Strukturaufklärung durch NMR erheblich beschleunigen können. Zu diesem Zweck konzentriert sich diese Arbeit auf bestimmte Messdaten aus der NMR, die so genannten *chemischen Verschiebungen*.

Chemische Verschiebungen werden standardmäßig zu Beginn einer Strukturauflösung bestimmt. Wie alle Labordaten können chemische Verschiebungen Fehler enthalten, die die Analyse erschweren, wenn nicht sogar unmöglich machen. Als erstes Resultat dieser Arbeit wird darum *CheckShift* präsentiert, eine Methode, die es ermöglich einen weit verbreiteten Fehler in chemischen Verschiebungsdaten automatisch zu korrigieren.

Das Hauptziel dieser Doktorarbeit ist es jedoch, strukturelle Informationen aus chemischen Verschiebungen zu extrahieren. Als erster Schritt in diese Richtung wurde *SimShift* entwickelt. SimShift ermöglicht es zum ersten Mal, strukturelle Ähnlichkeiten zwischen Proteinen basierend auf chemischen Verschiebungen zu identifizieren. Der Vergleich zu Methoden, die nur auf der Aminosäurensequenz basieren, zeigt die Überlegenheit des verschiebungsbasierten Ansatzes. Als eine natürliche Erweiterung des paarweisen Vergleichs von Proteinen wird darauf folgend *SimShiftDB* vorgestellt. Gegeben ein Protein, durchsucht SimShiftDB eine Datenbank bekannter Proteinstrukturen nach strukturell homologen Einträgen. Die Suche basiert hierbei nur auf der Aminosäuresequenz und den chemischen Verschiebungen des Proteins. Die detektierten Ähnlichkeiten werden zusätzlich nach statistischer Signifikanz bewertet.

Mit der *Chemical Shift Pipeline* wird schließlich das Hauptresultat der Dissertation vorgestellt. Durch die Kombination der automatischen Fehlerkorrektur (CheckShift) mit dem Datenbank-Suchalgorithmus (SimShiftDB), wird in 70% bis 80% der vorhergesagten strukturellen Ähnlichkeiten eine sehr hohe Qualität erreicht. Der Anteil der fehlerhaften Vorhersagen beträgt nur etwa 10%.

Summary

Nuclear magnetic resonance spectroscopy (NMR) is one of the most important methods for measuring the three-dimensional structure of biomolecules. Despite major progress in the NMR methodology, the solution of a protein structure is still a tedious and time-consuming task. The goal of this thesis is to develop bioinformatics methods which may strongly accelerate the NMR process. This work concentrates on a special type of measurements, the so-called *chemical shifts*.

Chemical shifts are routinely measured at the beginning of a structure resolution process. As all data from the laboratory, chemical shifts may be error-prone, which might complicate or even circumvent the use of this data. Therefore, as the first result of the thesis, we present *CheckShift*, a method which automatically corrects a frequent error in NMR chemical shift data.

However, the main goal of this thesis is the extraction of structural information hidden in chemical shifts. *SimShift* was developed as a first step in this direction. SimShift is the first approach to identify structural similarities between proteins based on chemical shifts. Compared to methods based on the amino acid sequence alone, SimShift shows its strength in detecting distant structural relationships. As a natural further development of the pairwise comparison of proteins, the SimShift algorithm is adapted for database searching. Given a protein, the improved algorithm, named SimShiftDB, searches a database of solved proteins for structurally homologue entries. The search is based only on the amino acid sequence and the associated chemical shifts. The detected similarities are additionally ranked based on calculations of statistical significance.

Finally, the *Chemical Shift Pipeline*, the main result of this work, is presented. By combining automatic chemical shift error correction (CheckShift) and the database search algorithm (SimShiftDB), it is possible to achieve very high quality in 70% to 80% of the similarities identified. Thereby, only about 10% of the predictions are in error.

Contents

1 Introduction

1.1 Overview

The main goal of Bioinformatics is to assist people working on biological problems with the solution of tasks that require a thorough understanding of computer science. One of the most prominent methods, which has nowadays become a standard tool for people working in molecular biology, is sequence comparison. With the size of the publicly available databases increasing, it became unfeasible to compare a newly derived sequence to a database of already solved and possibly functionally annotated sequences by hand. Therefore, tools were developed to search databases of millions of sequences automatically. To date, sequence comparison has a strong research background and many problems arising in this area have been solved.

Is the time for computational biology now over? Of course not, because there is more than just the protein sequence one might like to compare and investigate. What is strongly important in understanding the function of a protein is the protein's three-dimensional structure. It has been shown that there exist proteins with very low sequence similarity sharing a strongly similar three-dimensional structure (see for example [Pastore and Lesk, 1990]). This underlines the need for structural data as often sequence comparison alone is not sufficient. To date, it is not possible to calculate the protein structure reliably from the protein sequence alone. NMR spectroscopy is one of the most important methods for resolving a protein structure in the laboratory. However, solving a protein as a whole is a complicated task, requiring a series of complex experiments.

The work presented in this dissertation focuses on data which may be acquired fairly easily in a standard NMR experiment. This data are the so-called chemical shifts, which will be explained in section 2.2 in more detail. Chemical shifts are in general not sufficient to solve a protein structure and may contain various measurement errors. A thorough analysis of this data, however, can remove the need for additional experiments, thereby saving time and money.

Chapters 3 to 5 describe several computational methods which analyze chemical shift data, altogether reaching for the ultimate goal to be able to construct three dimensional models from chemical shift data alone. In Chapter 3 we present *CheckShift*, a method to correct a very common error in NMR chemical shift

data automatically. Chapter 4 describes and evaluates *SimShift*, a method which identifies structural similarities between proteins based on chemical shift data. It is shown that SimShift outperforms methods which work on amino acid sequence and/or secondary structure data alone. In Chapter 5, we extend and adapt this method for database searching. A statistical model is used to evaluate the significance of each similarity identified. *SimShiftDB* is able to search a whole database of protein structures for structural homologous, based on chemical shift information alone. The additional information hidden in chemical shifts proves to be useful to identify structural homologous for proteins where this is not possible using the amino acid sequence alone. Finally, Chapter 6 connects CheckShift and SimShiftDB thereby building the *Chemical Shift Pipeline*. We define a benchmark set and evaluate various parameter settings. The chemical shift pipeline achieves a correctness of up to 75% thereby covering 65% of the targets' residues.

1.2 Synopsis

NMR Spectroscopy is one of the most prominent methods for resolving protein structures on the atomic level. The solution of a protein structure, however, cannot be performed through a single experiment. Multiple sophisticated experiments with additional data analysis are necessary. In this work, we concentrated on a special type of data from NMR experiments, the so-called chemical shifts. Chemical shifts are easy to measure and are acquired by default at the beginning of the solution of a protein structure. In the beginning of our work, we asked the question: "Is it possible to identify structural similarities from chemical shifts, which may not be identified using the amino acid sequence alone?". To answer this question, we developed SimShift, a program which identifies structural similarities in proteins, solely based on chemical shifts. SimShift works as a two step algorithm. In the first step, locally similar regions of the two proteins are identified. In the second step of the algorithm, the best (legal) combination of a subset of these locally similar regions is calculated. To evaluate the performance of SimShift, we built a test set of pairs of (known) protein structures with associated chemical shifts. Now we compared SimShift to two state-of-the-art methods, which work solely on the amino acid sequence or the secondary structure of the respective protein. We were able to show that SimShift outperforms the other methods especially in the case of distant homologies. After having empirically proved, that it is possible to use the information hidden in chemical shifts we wanted to go one step further. Therefore, we developed a database search engine, called SimShiftDB, which searches a database of protein structures based on the chemical shifts measured for a target protein. We developed a scoring function which is able to reflect the probability that two residues with associated chemical shifts are part of the same three-dimensional structure.

Again a two step approach is used for each target-template pair which follows the same ideas as the SimShift approach. The underlying algorithms, however, were completely redesigned. For each target-template pair, SimShiftDB identifies structurally similar regions. If similar protein structures are found, these may be used to build a model for the target protein. These models may than be verified through a fairly simple NMR experiment. Building a model for a protein at this early stage of the NMR structure solution process is a very important task, as it saves time and money. To get high quality models, it is extremely important to evaluate the statistical significance of the similarities identified. Therefore, SimShiftDB calculates an E-Value for each similarity, which gives the number of results of equal or better quality expected to occur by chance. For the performance evaluation, we compare SimShiftDB to TALOS [Cornilescu et al., 1999], a widely used method which tries to infer protein backbone torsion angles from chemical shift data. As chemical shifts are error prone [Zhang et al., 2003], we used a small, but very reliable test set, given to us by collaborating researchers from NMR spectroscopy. SimShiftDB shows its strength especially in cases where TALOS gives erroneous predictions or no prediction at all. Therefore, SimShiftDB completes the range of existing methods which try to utilize chemical shifts for structure prediction. In parallel to the work on SimShiftDB, we realized that a very common error in NMR chemical shift data often hampers the quality of this data and, therefore, also the quality of any further data analysis. Chemical shifts are always given based on a reference compound to account for the different experimental conditions in different laboratories. Unfortunately, there are a great number of possible reference compounds available, and chemical shifts are often given without proper declaration of the reference compound used. Therefore, we developed a method, named CheckShift, which automatically corrects chemical shifts to a standard reference compound. CheckShift compares the target chemical shifts (which shall be corrected) to a database of chemical shifts which are reliably referenced according to the IUPAC [Markley et al., 1998] standard. Through this comparison it is possible to calculate the amount of correction necessary, which is then proposed as the re-referencing offset. The comparison to other methods which do automatically re-referencing showed that CheckShift has a significantly lower error rate. CheckShift is proposed as an error checker for newly deposited data by BMRB [Seavey et al., 1991], the most prominent public source for chemical shift data ("The Chemical Shift Reference Check" on `http://www.bmrb.wisc.edu/deposit`). In the most current work, we combined CheckShift and SimShiftDB, thereby building the Chemical Shift Pipeline. We built a benchmark set of protein structures with associated chemical shifts. For this reason, we tried to map each BMRB entry to a structure from the ASTRAL [Chandonia et al., 2004] database based on amino acid sequence homology. We were able to build a test set of 144 proteins structure with associated chemical shifts. Every protein in the test set was subsequently re-referenced using CheckShift. Then we evaluated differ-

ent parameter settings for SimShiftDB. To exclude trivial hits from the evaluation, template proteins which show high amino acid sequence similarity to the target protein were removed from the SimShiftDB results. Through the additional error correction and a certain parameter choice, the number of correct torsion angle predictions goes up to 75 percent, thereby achieving coverage of the targets' residues of 65 percent. SimShiftDB and CheckShift are available via `http://shifts.bio.ifi.lmu.de`.

2 Preliminaries

2.1 Introduction

In the beginning of this chapter basic information on the origin and measurement of chemical shifts is given. Afterwards we describe SHIFTX, which enables the user to back-calculate chemical shifts from the three-dimension protein structure. Following this section TALOS is introduced, a method which calculates protein backbone torsion angles based on chemical shift data. The short description of several Bioinformatics methods follows. These include methods for automatic calculation of secondary structure (PSIPRED, which is based on the amino acid sequence, CSI, which calculates secondary structure based on chemical shifts; STRIDE, which uses the three-dimension protein structure) and methods for calculating protein alignments based on the amino acid sequence (HHsearch), secondary structure (SSEA), as well as on the three-dimensional structure (CE). Another section is devoted to MaxSub, a score which evaluates the quality of an alignment based on the three-dimensional structure. Finally, an overview over the databases used for this work is given.

2.2 NMR Basics

Atoms possess a property called *spin*. An atom with spin $\frac{1}{2}$ has two possible stationary orientations in a magnetic field B_0. The spin (I) can take values $0, \frac{1}{2}, 1, \frac{3}{2}, \ldots$. In a magnetic field a nucleus of spin I has $2I+1$ possible orientations given by the magnetic quantum number m_I, which has values from $-I$ to I in integer steps. When a spin is inserted into a magnetic field, it has quantized[1] energy depending on its orientation in the field. The energy difference between the orientations is given by the equation

$$\Delta E = -\frac{\gamma h B_0}{2\pi} \ , \tag{2.1}$$

where h is Planck's constant and γ is the so-called *gyro-magnetic ratio* which determines the transition frequency of a nucleus in a given magnetic field. Atoms

[1] restricted to a certain set of values

in general have different energy levels. A transition between two energy levels can be caused by a photon with a certain frequency ν. In a magnetic field ν is dependent on the atom type (through the so-called *gyro-magnetic ratio* γ) and the strength of the magnetic field B_0: This frequency is proportional to the difference between the two energy levels, as expressed in the formula

$$\Delta E = h\nu \ . \tag{2.2}$$

Combining Equations 2.1 and 2.2, ν is obtained as

$$\nu = -\frac{\gamma B_0}{2\pi}$$

The *resonance* or *transition frequency* is therefore dependent on the atom type (through γ) **and** the applied magnetic field.

The state $m_I = \frac{1}{2}$ will be named α (parallel) and the state $m_I = -\frac{1}{2}$ β (anti-parallel).

The Boltzmann equation gives us information about the population of the two energy states α and β.

$$\frac{N_\beta}{N_\alpha} = e^{-\frac{\Delta E}{k_B T}}$$

As $\Delta E > 0, k_B = 1.38 * 10^{-23} > 0, T \geq 0$ the lower energy state (α) is populated more frequently at thermal equilibrium.

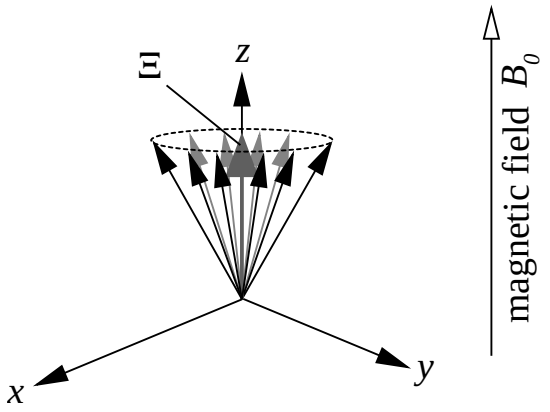


Figure 2.1: The magnetic moments of the nuclei precess uniformly distributed around the applied magnetic field B_0 (along the z-axis). The *bulk magnetization* Ξ of the individual nuclei is drawn as a thick arrow along the z-axis.

Consider the magnetic field B_0 to be in a three-dimensional coordinate system, where the direction of B_0 is along the z-axis. Each nucleus in the magnetic

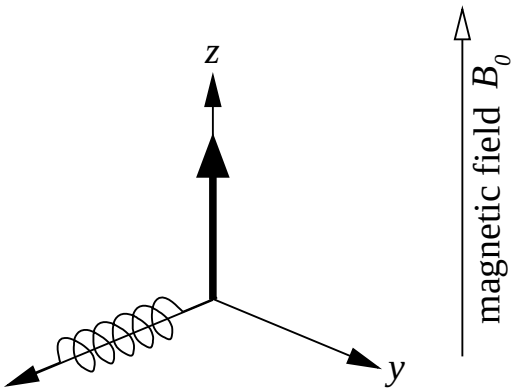


Figure 2.2: A coil produces an oscillating magnetic field perpendicular to the applied magnetic field B_0.

field has a certain *magnetic moment*. The magnetic moments of the nuclei precess around the z-axis with a certain frequency, called the *Larmor*-frequency (cf. Fig. 2.1). This frequency is equal to the transition frequency ν. The magnetic moments of the individual nuclei add up to a net macroscopic magnetization along the direction of the applied magnetic field B_0. This is called *bulk magnetization* of the sample and denoted by Ξ in Fig. 2.1.

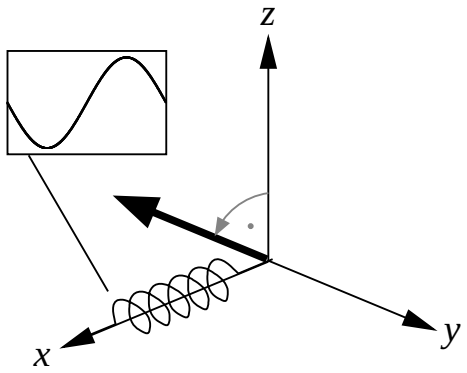


Figure 2.3: A coil produces an oscillating magnetic field perpendicular to the applied magnetic field B_0. This causes the bulk magnetization vector to be tilted away.

What is detected in the NMR experiment is the *precession* of the bulk magnetization vector around the z-axis (which equals the *Larmor*-frequency of the single nucleus). This precession starts when the bulk magnetization vector is tilted away from the z-axis by a magnetic pulse which oscillates along the x-axis (see Fig. 2.3). If the frequency of the magnetic pulse is equal or close to the *Larmor*-frequency, even a small magnetic field induced by the coil around the x-axis

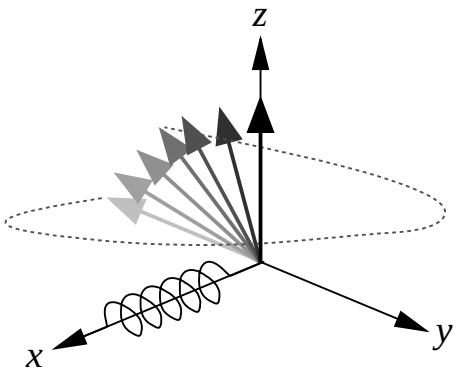


Figure 2.4: The bulk magnetization precesses around the z-axis back to its original position. This induces a current in the coil.

can tilt the bulk magnetization vector away from the z-axis. When rotating back to its original position, the magnetization vector induces a current in the coil in the xy-plane (see Fig. 2.4). This can be measured as the so-called **F**ree **I**nduction **D**ecay (FID). The FID is a function of the strength of the bulk magnetization along the x-axis over time. To convert this time domain function (FID) into a frequency domain function (spectrum), Fourier transformation is used.

Definition 1. *The* **chemical shift** *of an atom is its resonance frequency relative to some standard reference compound S. It is defined as*

$$\delta = \frac{\nu - \nu_S}{\nu_S} \ ,$$

where ν_S is the frequency of the standard *S and ν is the frequency represented by the chemical shift value. As these values tend to be very small, they are usually given in ppm (parts per million).*

The chemical shift for a certain nucleus is influenced by its environment. This is due to the fact that surrounding electron clouds induce a local magnetic field, which adds to (or subtracts from) the applied field. Therefore, the effective field acting upon the nucleus is different from the applied magnetic field B_0. As the chemical shift is dependent on the resonance frequency, and the resonance frequency is in turn dependent on the magnetic field acting on the nucleus, this results in different chemical shifts for different environments. This is why the shift values are helpful for resolving (protein) structures with NMR. In a protein the chemical shift of an atom is influenced by the type of amino acid it is part of *and* the electron clouds of atoms which are close in space due to the tertiary structure of a protein. The influence of the amino acid type may be removed by subtracting the so-called *random coil shifts*, which give an average value for the chemical shift of a specific atom in a certain amino acid. These normalized chemical shift values are then called *secondary shifts*.

For more information on chemical shifts and NMR in general, we refer the reader to [Levitt, 2001].

2.3 SHIFTX

SHIFTX [Neal et al., 2003] is a computer program which rapidly and accurately calculates the diamagnetic ^{1}H, ^{13}C, and ^{15}N shifts of both backbone and side chain atoms in proteins. To calculate the chemical shifts from the structure, SHIFTX uses a hybrid approach, combining classical equations with chemical shift hypersurfaces, derived from a database of three-dimensional structures with associated chemical shifts. The chemical shift for a specific atom (δ) is calculated as the sum of several components:

$$\delta = \delta_{\mathrm{C}} + \delta_{\mathrm{RC}} + \delta_{\mathrm{EF}} + \delta_{\mathrm{HB}} + \delta_{\mathrm{HS}}, \tag{2.3}$$

where δ_{C} is the random coil shift as given by Wishart et al. [1995a], δ_{RC} is the ring current shift, δ_{EF} is the electric field contribution, δ_{HB} is the hydrogen bond contribution and δ_{HS} is the contribution of the chemical shift hypersurfaces. The SHIFTX program performs the following steps sequentially:

1. Check and calculate the positions of the hydrogen atoms. (For details, see [Neal et al., 2003], section *Hydrogen placement*.)
2. Calculate ring current, electric field, and hydrogen bond contributions.
3. Calculate chemical shift hypersurface contributions.
4. Sum all contributions to calculate the predicted chemical shift.

2.3.1 Ring Current Effects

Aromatic rings have a strong influence on the chemical shifts of nearby nuclei. SHIFTX first calculates a list of rings and a list of atoms, which may be influenced. Subsequently, the influence on each chemical shift for each susceptible atom is calculated using the methods by Haigh and Mallion [1972]. The ring current contribution is calculated as

$$\delta_{\mathrm{RC}} = \mathbf{G}*\mathbf{I}*\mathbf{F}, \tag{2.4}$$

where $\mathbf{G}$ is a geometrical factor, $\mathbf{I}$ is a ring specific intensity, and $\mathbf{F}$ is a atom specific constant. The parameter $\mathbf{G}$ is equivalent to the parameter K_i' described in [Haigh and Mallion, 1972]. $\mathbf{I}$ and $\mathbf{F}$ were determined empirically using the training database (see tables 2.1 and 2.2 for the chosen values).

Table 2.1: Empirically optimized values for the ring specific intensity (**I**).

Residue	I	Ring Atoms
F	1.05	CG-CD2-CE2-CZ-CE1-CD1
Y	0.92	CG-CD2-CE2-CZ-CE1-CD1
W	1.04	CD2-CE3-CZ3-CH2-CZ2-CE2
W	0.90	CG-CD2-CE2-NE1-CD1
H	0.43	CG-ND1-CE1-NE2-CD2

Table 2.2: Empirically optimized values for the atom specific constant (**F**).

Atom	F
HN	$7.06 * 10^{-6}$
Other H	$5.13 * 10^{-6}$
CA	$1.50 * 10^{-6}$
CB, CO, N	$1.00 * 10^{-6}$

2.3.2 Electric Field Effects

Alpha carbons and all hydrogens are influenced by electrostatic effects. To calculate the influence on the chemical shifts of the respective atoms, the method by Buckingham [1960] is used. Atoms are therefore classified as *target* (influenced through electrostatic effects) and *source* atoms (reason for the electrostatic effects). The calculation also requires the coordinates of the *partner* atoms for each target atom. Partner atoms are bonded to the target atom. A list of partner atoms for each target is given in Table A.1 in Appendix A. All source-target combination within 3.0Å are analyzed, given the following constraints hold:

- Source and target atom are **not** part of the same residue or adjacent residue.
- If the target is atom HN, source must not be O.
- Solvent atoms do **not** act as sources.

The effect on the chemical shift can then be calculated as:

$$\delta_{\mathrm{EF}} = \frac{10^{10} * q * \cos\theta}{d^2}, \tag{2.5}$$

where q is the source charge as according to table 2.3, θ is the source-target-partner angle and d is the distance from source to target in Ångstrom.

Table 2.3: Partial charges for each source atom ([esu]).

Atom(s)	Charge
O, OD, OE	$-0.9612 * 10^{-10}$
C	$1.3937 * 10^{-10}$
N	$0.7209 * 10^{-10}$

2.3.3 Hydrogen Bond Effects

It was found that including hydrogen bond effects into the calculation improves the performance of SHIFTX. To calculate the influence of hydrogen bonds on the chemical shifts, a list of donor-acceptor pairs is compiled. The following constraints have to be fulfilled for all pairs:

- The donor and the acceptor must be on different residues.
- If the acceptor is a solvent oxygen, the donor must be H_α.
- The oxygen-hydrogen distance must be less than an empirically derived cutoff (3.50Å for HN and 2.77Å for H_α).
- The vector between the N-H bond and the C=O bond must be 90° or more, when the vectors are translated such that N and C occupy the same point in space.

SHIFTX now builds a list of possible hydrogen bonds. Subsequently, this list is processed, such that only the strongest hydrogen bond for each donor is kept in the list. Then the following formula (see [Wagner et al., 1983, Wishart et al., 1991]) is used to calculate the effects on the chemical shifts of the respective atoms:

$$\delta_{\mathrm{HB}} = \frac{0.75}{r^3} - 0.99. \tag{2.6}$$

Thereby r is the distance between donor and acceptor. The parameters (0.75 and 0.99) were optimized on the SHIFTX training set. The formula may be applied to bonds of length between 1.5 to 3.5Å. To account for the (more infrequently occurring) case in which HA protons act as hydrogen bond donors, a second equation is used:

$$\delta_{\mathrm{HB}} = \frac{15.56}{r^3} - 0.67. \tag{2.7}$$

For cases in which the distance r is greater than 2.61 or less than 2.27, r is set to a fixed value of 2.61 or 2.27, respectively.

2.3.4 Empirical Chemical Shift Hypersurfaces

SHIFTX uses a set of so-called chemical shifts hypersurfaces (two-dimensional arrays) which give chemical shift corrections for a variety of structural parameters, whose influence is not taken into account using the classical equations described above. To identify the necessary corrections and the parameters having the strongest influence on the chemical shift, a data mining procedure was applied (details on the procedure are not given in the publication). The analyzed parameters are listed in Table 2 in [Neal et al., 2003]. Every two-parameter-combination was tested and the ones showing a strong need for chemical shift correction were chosen. The parameter combinations (and the associated chemical shift hypersurfaces) finally used are described in the supplementary material of [Neal et al., 2003]. The influence of all applicable hypersurfaces for a given atom are then summed to calculate δ_{HS}.

It has to be noted that the description given here is a strong simplification of the procedure actually used, as not even in the original SHIFTX publication the full detail of the chemical shifts hypersurface calculations are given.

2.4 TALOS

TALOS [Cornilescu et al., 1999] is one of the most widely used tools for NMR spectroscopists. The aim of TALOS is to identify protein backbone angles by comparing a set of newly derived chemical shifts to a database of chemical shifts which are associated with an already resolved protein structure.

TALOS moves a sliding window of length three over the residue sequence of the target protein. Each triple from the target is then compared to each triplet in the template database using an optimized scoring function. The scoring function is defined as

$$\begin{aligned} S[t_i][s_j] = \sum_{k=-1}^{1} \Big\{ & \mu_M[k] * M[t_{i+k}][s_{j+k}] + \mu_{C_\alpha}[k] * (\overline{\delta}_{C_\alpha}[t_{i+k}] - \overline{\delta}_{C_\alpha}[s_{j+k}]) \\ & + \mu_{C_\beta}[k] * (\overline{\delta}_{C_\beta}[t_{i+k}] - \overline{\delta}_{C_\beta}[s_{j+k}]) + \mu_{C'}[k] * (\overline{\delta}'_{C}[t_{i+k}] - \overline{\delta}'_{C}[s_{j+k}]) \\ & + \mu_N[k] * (\overline{\delta}_N[t_{i+k}] - \overline{\delta}_N[s_{j+k}]) + \mu_{H_\alpha}[k] * (\overline{\delta}_{H_\alpha}[t_{i+k}] - \overline{\delta}_{H_\alpha}[s_{j+k}]) \Big\} \end{aligned} \tag{2.8}$$

where $\overline{\delta}_{C_\alpha}, \overline{\delta}_{C_\beta}, \overline{\delta}'_C, \overline{\delta}_N$, and $\overline{\delta}_{H_\alpha}$ are the secondary shifts for the respective atoms.

This scoring function compares the residue types (using the substitution matrix M, see [Cornilescu et al., 1999] for details), but also takes into account the differences between C_α, C_β, C', N and H_α secondary shifts (if available in the target).

Table 2.4: Weighting factors for scoring function

k	μ_M	μ_{C_α}	μ_{C_β}	$\mu_{C'}$	μ_N	μ_{H_α}
-1	0.739	0.7213	0.7624	1.1455	0.1596	14.665
0	1.478	0.9857	0.9092	1.2051	0.1752	17.539
1	0.739	0.7178	0.6990	1.0422	0.1972	15.251

Table 2.5: RMSD values of the secondary shifts for C_α, C_β, C', N and H_α

C_α	C_β	C'	N	H_α
2.40	2.01	2.02	4.56	0.51

To find the best weighting factors for the chemical shift differences (μ_{C_α}, μ_{C_β}, $\mu_{C'}$, μ_N, μ_{H_α}), the (original) template database was searched with 183 residue triples. For each target triplet, the RMSD and the standard deviation of all database triples, where the difference between the target's and template's Φ and Ψ angles is less than 15°, was calculated. Then the average of the RMSD divided by the standard deviation is used as the weighting factor for the respective atom (see Table 2.4).

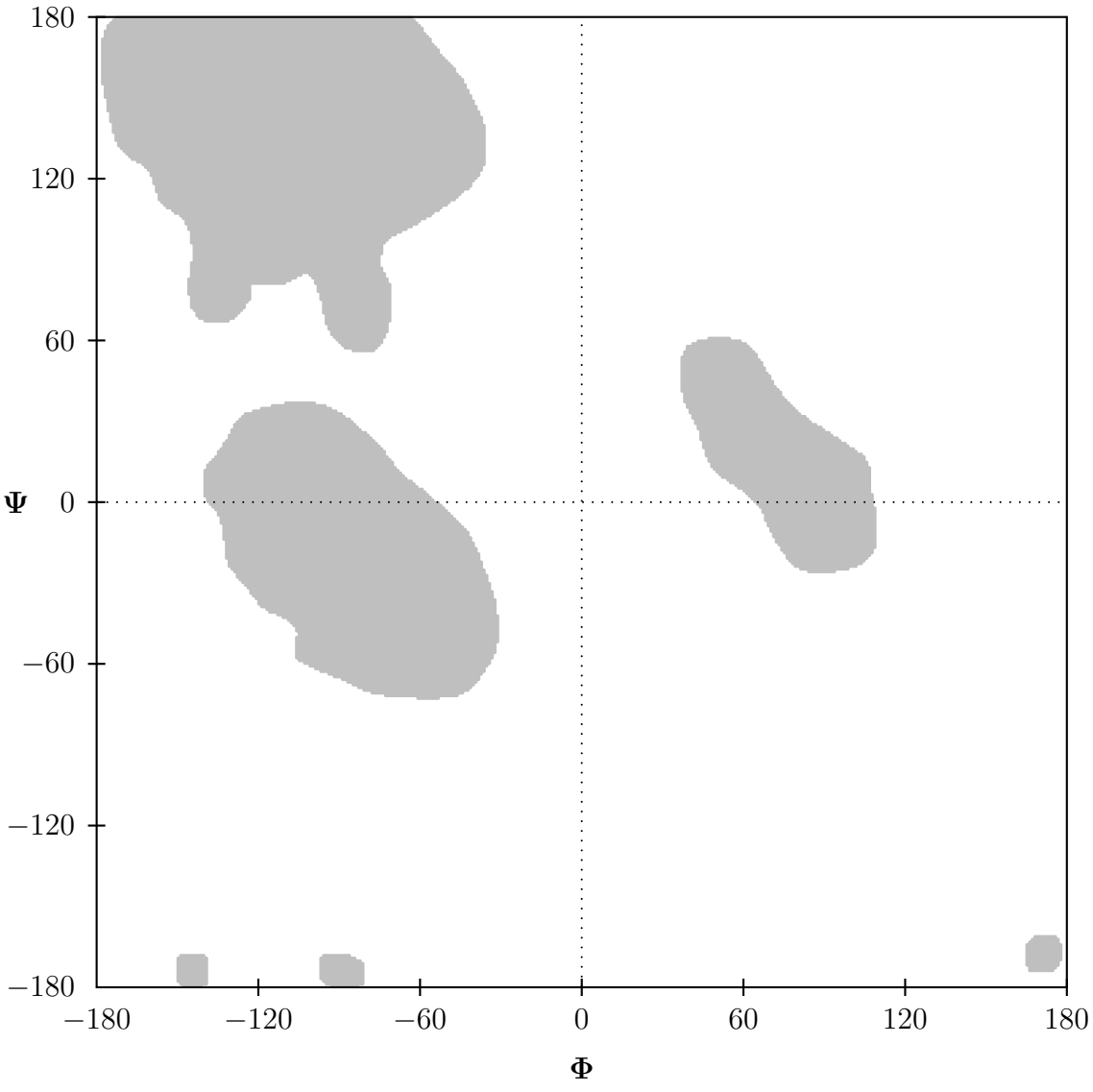
$\mu_M$was optimized empirically, taking those factors which minimize the number of erroneous prediction. More detailed information on how $\mu_M$was optimized (test set, value range) are not available in [Cornilescu et al., 1999].

If chemical shifts are missing in the target, the value is set to 1.5 times the RMSD of the corresponding secondary shift (see Table 2.5). Therefore, missing chemical shift values should not contribute to a good score. To evaluate the quality of the predictions, the 10 best matching triplets (according to $S[t_i][s_j]$) for every target triplet are analyzed. The following classifications are possible in the standard version:

Good: Either 9 out of the 10 triplets have (Φ, Ψ) angles in the same region of the Ramachandran map (see Figure 2.5) and none of the center residues in the 10 triplets has a positive Φ angle, or 9 triplets are situated in the positive Φ angle range.

New: All residues that do not fulfill the constraints for **Good**.

Figure 2.5: Ramachandran map used for TALOS torsion angle classification.



2.5 PSIPRED

PSIPRED [Jones, 1999] predicts the secondary structure of a protein based on the amino acid sequence alone. This is accomplished in the following way:

- A PSI-BLAST [Altschul et al., 1997] search using the target sequence is performed against a non-redundant database, which is beforehand filtered for low complexity regions. Thereby, the number of iterations for PSI-BLAST is set to 3.

- The calculated position specific scoring matrix is subsequently used as an input to a neuronal network which gives the three state secondary structure prediction as an output.

The neural network uses a standard feed-forward architecture and was trained by applying a back-propagation procedure Rumelhart et al. [1986]. In secondary structure prediction it is a common procedure to include neighboring residues, when trying to predict the central residue's structure. Based on a small benchmark set (16 targets from the CASP2 experiment Moult et al. [1997]), a window of 15 residues (7 residues on either side of the target residue) was selected to be optimal in performance. The whole network consists of two parts, the first part being built of 315 input units (21 input units for every residue in the window), 75 hidden units and 45 output units. Thereby, the 21st residue is used as an indicator that the window spans a chain terminus. 45 outputs are then fed to the second part of the network consisting of 60 input units (4 secondary structure states times 15 residues in the window, the 4th secondary structure state being used as the 21st amino acid before), 60 hidden units and finally 3 outputs for the prediction of the central residue.

On the benchmark set presented by [Jones, 1999], PSIPRED achieves an average Q3 score (average accuracy for the three secondary structure states) between 76.5% and 78.3%. According to an evaluation by the EVA server [Rost and Eyrich, 2001] conducted on March 27, 2007, PSIPRED still ranks at first place concerning the comparison of average scores and under the best 4 methods available as according to a pairwise comparison.

2.6 Chemical Shift Index (CSI)

The chemical shift index (CSI) [Wishart et al., 1992] is a simple and very robust method to calculate the secondary structure of a protein based on $^1H_\alpha$ chemical shift values. The calculation works in 3 steps:

1. The chemical shift values are normalized using amino acid specific random coil shifts.

2. The resulting secondary shift values are converted to discrete values (-1,0,1), by setting all values less than -0.1 to -1, more than 0.1 to 1 and all between to zero. Thereby -1 stands for potential helix residues, 0 for coil residues and 1 for residues being part of a strand.

3. Several heuristics are applied, including a minimum length restriction for secondary structure element, thereby producing the final output.

CSI was enhanced to work also using ^{13}C chemical shifts [Wishart and Sykes, 1994]. The ^{13}C-method works exactly as the one described for $^{1}H_{\alpha}$ chemical shift values. The zero-range, however, is defined as [-0.7,0.7]. Using a consensus prediction of $^{1}H_{\alpha}$ and ^{13}C chemical shifts, CSI is able to achieve an accuracy in excess of 92%.

2.7 Structural Identification (STRIDE)

STRIDE [Frishman and Argos, 1995] calculates secondary structure from the three dimensional structure of a protein. These calculations are based on hydrogen bonding patterns as well as on backbone torsion angles. STRIDE calculates an empirical energy for hydrogen bonds and probabilities for torsion angles being part of an alpha helix or a beta strand. For the assignment of a certain secondary structure type, the hydrogen bond energies and probabilities are combined and compared to a cutoff. The weighting factors and the cutoff values were optimized using a hand curated dataset of protein structure with secondary structures defined by the experimentalists. STRIDE is shown to yield assignments closer to those given in PDB for nearly twice as many structures as the most famous methods for secondary structure calculation, namely DSSP [Kabsch and Sander, 1983]. Assignments made by STRIDE are in general in agreement with DSSP (the maximal difference in percent of correctly assigned residues does not exceed 14%). Based on these facts, we consider STRIDE as the better alternative and, therefore, used it throughout this work when secondary structure had to be assigned.

2.8 HHsearch

HHsearch [Söding, 2005] is a state-of-the-art method to align amino acid sequence through the comparison of two hidden markov models. When compared to other sequence based methods (PSI-BLAST [Altschul et al., 1997], HMMER [Eddy,

1998]), and profile-profile comparison tools (PROF_SIM [Yona and Levitt, 2002], COMPASS [Sadreyev and Grishin, 2003]), it is shown that HHsearch outperforms the other methods both in the detection of homologues and in alignment quality. Additionally, HHsearch is able to build alignments based on both amino acid sequence and secondary structure. This is especially useful for this work, as it is possible to calculate secondary structure quite reliably from chemical shift data [Wishart et al., 1992].

2.9 Secondary Structure Element Alignment (SSEA)

SSEA [Fontana et al., 2005] calculates protein alignments based on secondary structure alone. Thereby, the secondary structure elements (continues stretches of a certain secondary structure) are derived from both secondary structure strings to be aligned. The secondary structure has to be classified to one of three states, being either helix, strand, or coil. For the alignment, a simple scoring function is applied:

- Matching secondary structure elements (helix to helix, strand to strand, or coil to coil) are scored by the length of the smaller element.
- Mismatches (helix to strand) are not scored at all.
- Structure to coil matches are scored half the length of the shorted segment.

Additionally, the user may choose to calculate either a global or a local alignment.

2.10 Combinatorial Extension (CE)

CE [Shindyalov and Bourne, 1998] calculates protein structure alignments based on the combination of so-called aligned fragment pairs (AFPs). These AFPs are identified by searching for strong local similarities in the two structures. The complete alignment is constructed in three steps:

1. Potential AFPs (of fixed length) are filtered by comparing all residue-residue distances in the two protein fragments (see the description of the D_{Single} score below). This leaves the algorithms with a set of high quality AFPs, which are subsequently used as possible starting points of the alignment path.

2. Then the algorithm tries to identify the best consistent combination of AFPs. Every AFP from the last step is considered as a possible starting point for the final alignment path. To evaluate if an AFP shall be added to an existing path, an independent set of residue-residue distances between the new AFP and all the AFPs already in the path is compared. This is done in a pairwise fashion using the D_{Pair} score described below. If the average difference between the intra-protein distances satisfies an empirically derived cutoff, the AFP is added to the path.

3. For alignment paths of sufficient quality, an optimization procedure is applied. This last step is based on minimizing the root mean square deviation of aligned residues in the superimposed three-dimensional structures of the two proteins.

CE compares well to other structure alignment methods such as DALI or VAST (see Tables 5 and 6 in [Shindyalov and Bourne, 1998]). Additionally, CE was used in the CASP5 [Kinch et al., 2003] and CASP6 [Wang et al., 2005a] experiment for performance evaluation.

2.10.1 Distance Scores

The D_{Single} score is used to evaluate the quality of a single AFP. It is defined as follows,

$$D_{\text{Single}} = \frac{1}{l^2} \sum_{i=0}^{l-1} \sum_{j=0}^{l-1} |d^{\mathbf{A}}_{s^{\mathbf{A}}+i,s^{\mathbf{A}}+j} - d^{\mathbf{B}}_{s^{\mathbf{B}}+i,s^{\mathbf{B}}+j}|, \tag{2.9}$$

where l is the length of the AFP, $s^{\mathbf{A}}$ and $s^{\mathbf{B}}$ are its starting points in protein $\mathbf{A}$ and $\mathbf{B}$ respectively, and $d^{\mathbf{A}}_{i,j}, d^{\mathbf{B}}_{i,j}$ are the distances between residues i and j in the two proteins.

To evaluate the quality of the combination of two AFPs, the score D_{Pair} is used. It is defined in the following way,

$$\begin{aligned} D_{\text{Pair}} = & \frac{1}{l}\Big(|d^{\mathbf{A}}_{s_1^{\mathbf{A}},s_2^{\mathbf{A}}} - d^{\mathbf{B}}_{s_1^{\mathbf{B}},s_2^{\mathbf{B}}}| + |d^{\mathbf{A}}_{s_1^{\mathbf{A}}+l-1,s_2^{\mathbf{A}}+l-1} - d^{\mathbf{B}}_{s_1^{\mathbf{B}}+l-1,s_2^{\mathbf{B}}+l-1}| \\ & + \sum_{i=1}^{l-2} |d^{\mathbf{A}}_{s_1^{\mathbf{A}}+i,s_2^{\mathbf{A}}-i-1+l} - d^{\mathbf{B}}_{s_1^{\mathbf{B}}+i,s_2^{\mathbf{B}}-i-1+l}| \Big), \end{aligned} \tag{2.10}$$

where the meaning of the variables is the same as for the single score, despite of the starting points $s_1^{\mathbf{A}}, s_2^{\mathbf{A}}, s_1^{\mathbf{B}}, s_2^{\mathbf{B}}$, which are now not only associated to a certain protein, however, the subscripts additionally refer to either the first or the second AFP.

2.11 MaxSub

The MaxSub score [Siew et al., 2000] is an alignment dependent measure to evaluate the quality of a certain protein-protein alignment based on the quality of the superposition of the aligned residues. It is defined as

$$S = \frac{\sum_i \frac{1}{1+(\frac{d_i}{d})^2}}{n} \tag{2.11}$$

where d_i is the distance between the ith pair of equivalent residues after superposition, d is a cutoff parameter (the authors propose a value of 3.5 Å), and n is the maximal number of residues which could be aligned.

The MaxSub score is chosen for the evaluations presented in Chapter 4 due to the following reasons:

- It is a simple and robust score, which represents a good tradeoff between the number of aligned residues and the quality of the alignment.
- MaxSub only requires a single parameter (d) and it was shown by Siew et al. [2000] that the score is stable with respect to this parameter in a range from 2Å to 7Å.

MaxSub proves its discriminative power on the CASP3 targets as there is good agreement between the human based evaluation of the better models and the evaluation through the MaxSub score. MaxSub is also used for prediction evaluation in all CAFASP experiments starting from CAFASP2 [Fischer et al., 2001].

2.12 Databases

Several publicly available databases are used in this work. The spectrum reaches from protein structure databases (ASTRAL), and databases containing pairwise (DALI Database) and multiple (DMAPS, S4) structural alignments, to the the main repository for chemical shift data (BMRB). A short description of each database used is given in the following.

ASTRAL: The ASTRAL database [Chandonia et al., 2004] contains a filtered subset of protein structures from the PDB [Berman et al., 2000]. The filtering process removes redundancy from the PDB set of structures and additionally splits the structures according to the SCOP [Murzin et al., 1995] classification. In the process of removing redundancy, representatives have to be chosen for sets of equivalent structures. Thereby, the focus is laid on choosing the structure of highest quality available.

DALI Database: The DALI Database [Holm and Sander, 1996] is constructed from an all against all comparison of all entries in PDB. It contains the corresponding structural alignments which are calculated using the DALI search engine. Here, we use the alignments in this database as a basis for calculating chemical shift substitution matrices.

DMAPS: DMAPS [Guda et al., 2006] contains multiple structural alignments for several sets of protein families. The alignments are calculated using the CE-MC algorithm [Guda et al., 2004]. In this work, we use two of the available alignment sets. The first set contains multiple structural alignments for each SCOP domain family. The second set defines families through a certain extent of structural similarity identified by the CE algorithm. Therefore, an all-against-all comparison of all structures in PDB is performed and neighbors with a z-score > 4.0 and and RMSD < 3.0Å are assembled into clusters of common substructures. For each of those clusters multiple structure alignments are calculated.

S4: S4 [Casbon and Saqi, 2005] is an automatically generated database. It contains multiple protein structure alignments for each superfamily as defined by the SCOP classification. Thereby, structural domains may not share more than 40% sequence identity with another member in the superfamily to be included in the alignment.

BMRB: The BMRB [Seavey et al., 1991] is the main repository for data from NMR spectroscopy, measured from proteins, peptides and other biomolecules. Each entry in the database is associated to a specific protein and contains assigned chemical shifts, as well as additional data from the corresponding NMR experiments.

3 CheckShift

3.1 Introduction

The most common approach to extracting structural information from protein chemical shifts is to compare the shifts of the target protein to a database of reference shifts. This has been applied to direct refinement of protein structures [Schwieters et al., 2003], prediction of protein secondary structure [Wishart et al., 1992, Wang and Jardetzky, 2002], inference of protein backbone angles [Cornilescu et al., 1999, Neal et al., 2006], structure validation [Oldfield, 1995], and detection of structural similarities in proteins [Ginzinger and Fischer, 2006, Ginzinger et al., 2007]. In all of these methods, the quality of the database is crucial to the outcome, in terms of its size, the accuracy of the component structures, and consistent referencing of chemical shifts. The last factor is perhaps a larger obstacle than it may first appear, due to the number of different referencing compounds and methods in current use. Even with detailed information on the method, re-referencing of shifts to a single standard is difficult. In practice, incomplete or inconsistent annotation in the main repository, the Biological Magnetic Resonance Database (BMRB) [Seavey et al., 1991], often makes this impossible, and cases where re-referencing is necessary can be difficult to detect. Often the magnitude of referencing errors is of the same order as structure-dependent secondary shifts, and thus all data must be checked for accurate referencing before use [Zhang et al., 2003].

Several existing programs are capable of re-referencing chemical shifts using expectation values calculated on a residue-by-residue basis either from high-resolution structures [Neal et al., 2003, Zhang et al., 2003] or secondary structure predictions based on correctly referenced $^1H_\alpha$ shifts [Wang and Wishart, 2005]. Here we present a method for automatically re-referencing chemical shift data, named *CheckShift*, which takes the alternative approach of comparing the global chemical shift distribution of the target protein to a reference distribution. In addition to the chemical shift values, CheckShift requires only an estimate of the overall proportion of residues in β-sheet and α-helix secondary structures, a quantity that can be reliability predicted from primary sequence using PSIPRED [Jones, 1999]. CheckShift minimizes the difference between the distributions' density functions. Due to this modus operandi, CheckShift is in-

sensitive to outlying values. We show here that CheckShift is very accurate and compares well to other structure independent methods.

3.2 The CheckShift Algorithm

The following steps are performed to calculate the re-referencing offset for each atom type of a set of target chemical shifts. Each step will be discussed in detail below.

1. **Preparation of reference density functions:** Secondary shift density functions from correctly referenced data sets are prepared as a reference. This step has to be performed only once.
2. **Calculation of similarity:** The reference density functions are compared to the density function of the secondary shifts in the target data set.
3. **Re-referencing of data sets:** The previous step is iterated while changing the re-referencing offset to find the best agreement of the target and reference. The offset that minimizes the difference between the two density functions is suggested as re-referencing offset.

3.2.1 Preparation of Reference Density Functions

We have used all ^{13}C', $^{13}C_{\alpha}$, $^{13}C_{\beta}$ and ^{15}N chemical shifts which are included in the TALOS reference database (78 proteins, referenced to DSS and liquid ammonia) to prepare the reference density function. The TALOS database was chosen as it is well curated and therefore of very high quality, as opposed to other sources of chemical shift data. Chemical shifts from cysteine residues are excluded as they strongly depend on the oxidation state of each residue, which is a structure dependent feature that cannot be predicted using sequence information alone. Although structures are available for all entries from TALOS and, thus, cysteine oxidation states are known, this is not necessarily the case for the target chemical shifts.

Subsequently, the secondary shifts for all chemical shifts from the remaining 19 amino acids are derived by subtracting the amino acid-specific random coil shifts as given by Zhang et al. [2003]. The secondary structure associated with each chemical shift is calculated from the corresponding protein structure using STRIDE [Heinig and Frishman, 2004]. The 5 letter code given by STRIDE is converted to a three letter code as follows: **G**, **I** and **H** are translated to helix (**H**), **B** and **E** are defined as sheet (**S**) and all others are set to coil (**C**). Therefore, the secondary shifts can be classified according to their secondary structure. This

gives rise to three separate secondary shift density functions for each atom type (see Figure 3.1). Please note that the number of shifts in each distribution is different, leading to a prior probability $\rho = (\rho_H, \rho_S, \rho_C)$ for each of the three secondary structure states.

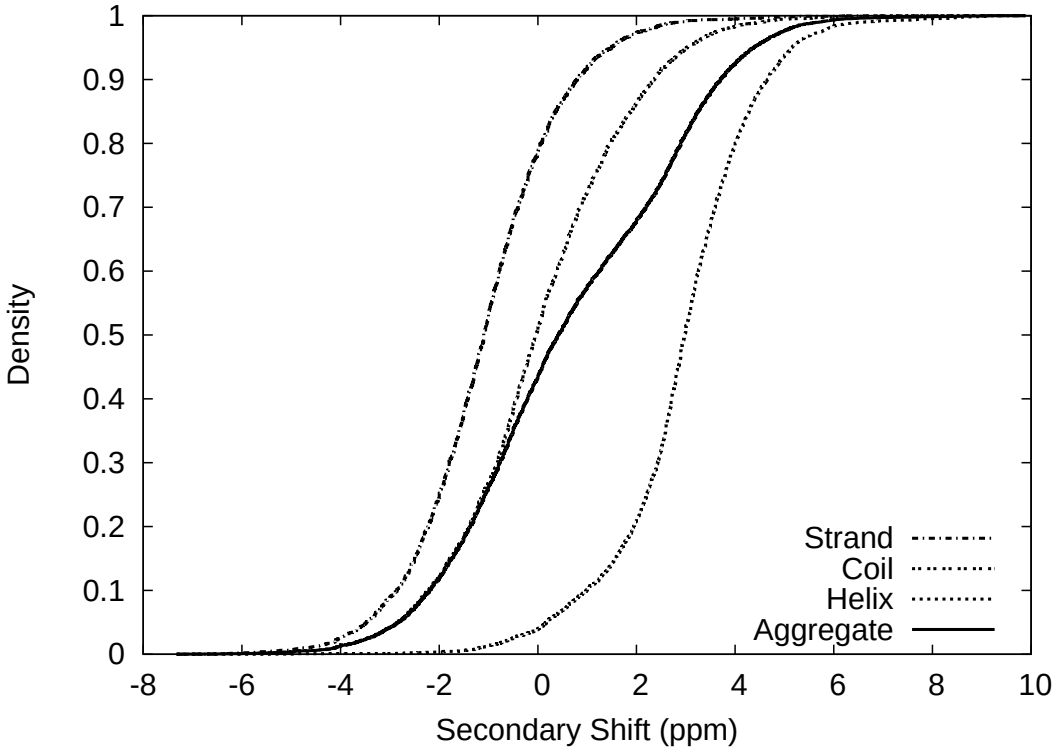


Figure 3.1: Density function of $^{13}C_\alpha$ secondary shifts from TALOS, used as expectation for secondary shifts of correctly TSP-referenced data sets. The density functions for each of the three secondary structures states (Sheet, Coil, Helix) are shown together with the total density function (Aggregate).

3.2.2 Calculation of Similarity

When predicting the re-referencing offset for each atom type of a target, the three secondary structure dependent density functions serve as the reference. These are based on the empirical chemical shifts of proteins, which are referenced according to the IUPAC standard. Target chemical shifts, which are given in the standardized way, are expected to have a similar density function as the reference. On the other hand, if the density functions are found to be shifted, this is an indicator of non-standard referencing or a referencing error.

For the comparison, secondary shifts are derived from the target's chemical shifts, except for cysteine. Subsequently, PSIPRED [Jones, 1999] is used to predict the secondary structure of the target sequence. This is due to the fact that three dimensional structures are not always available, and thus neither a mapping nor a defined secondary structure can be derived. While PSIPRED in general gives good predictions of secondary structures, this prediction is not used to split the secondary shifts of the target according to the secondary structure, but only to calculate the ratio $\sigma = (\sigma_H, \sigma_S, \sigma_C)$ of the three secondary structure states relative to each other. Later, for each of the three secondary structure states $sec \in (H, S, C)$, the respective secondary structure dependent reference density function from TALOS with a prior ρ_{sec} is scaled by σ_{sec}/ρ_{sec} to have the same ratio σ as the target protein before combining and comparing them to the target's combined density function. Please note the difference between the two density functions in Figure 3.2 for an illustration of this approach. This takes into account that proteins can have a very different secondary structure content, having a related ratio σ that is not necessarily equal to the prior ρ from TALOS. Consequently, this leads to different expected secondary shift density functions.

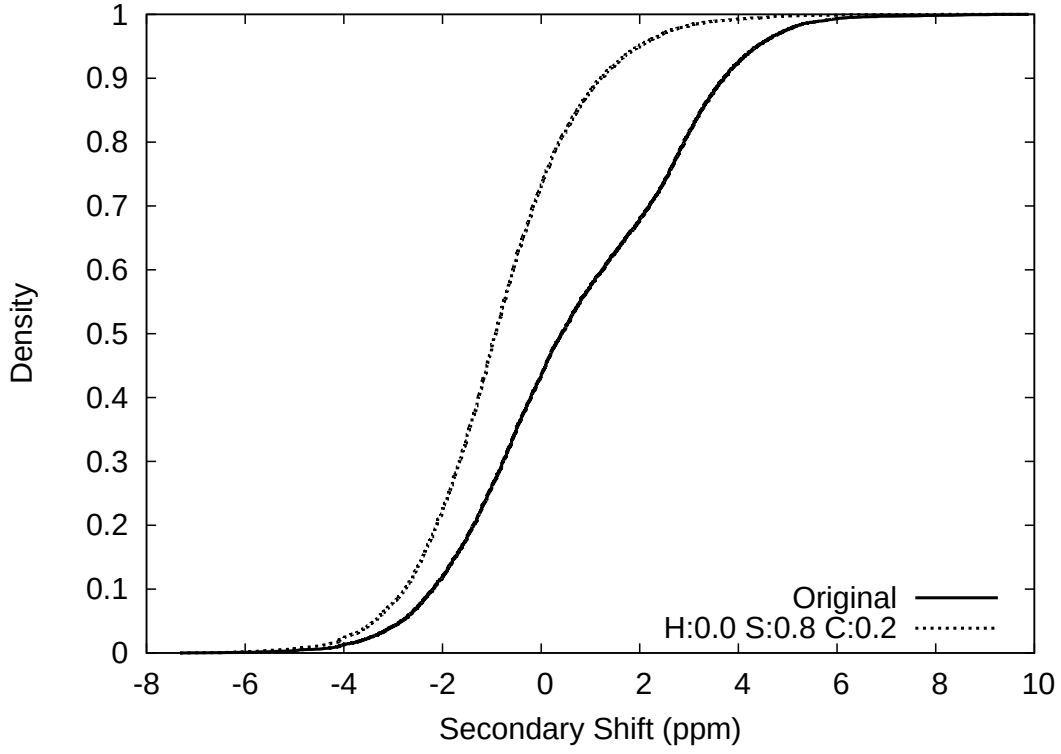


Figure 3.2: The density function of the $^{13}C_\alpha$ secondary shifts from the TALOS data set together with the adjusted density function for a protein with 80% β-content, i.e. $\sigma = (0.0\%, 0.8\%, 0.2\%)$.

This approach avoids a wrong assignment of secondary shifts to a specific secondary structure, which would occur by splitting secondary shifts based on the secondary structure prediction. Wrong prediction of secondary structure would then result in inferior secondary shift density functions. Consequently, checking consistency to the reference distributions would be more difficult and error-prone. While PSIPRED makes correct predictions with a rate of about 83%, its strength is to correctly predict the overall architecture of whole secondary structure elements. However, the *exact* positions of those elements is not always predicted correctly, and may vary by a few residues. Therefore, using only the information about the overall secondary structure architecture (i.e., secondary structure content), combining the three scaled density functions, and comparing the two density functions as described above, should be more accurate than using the information in a residue specific way.

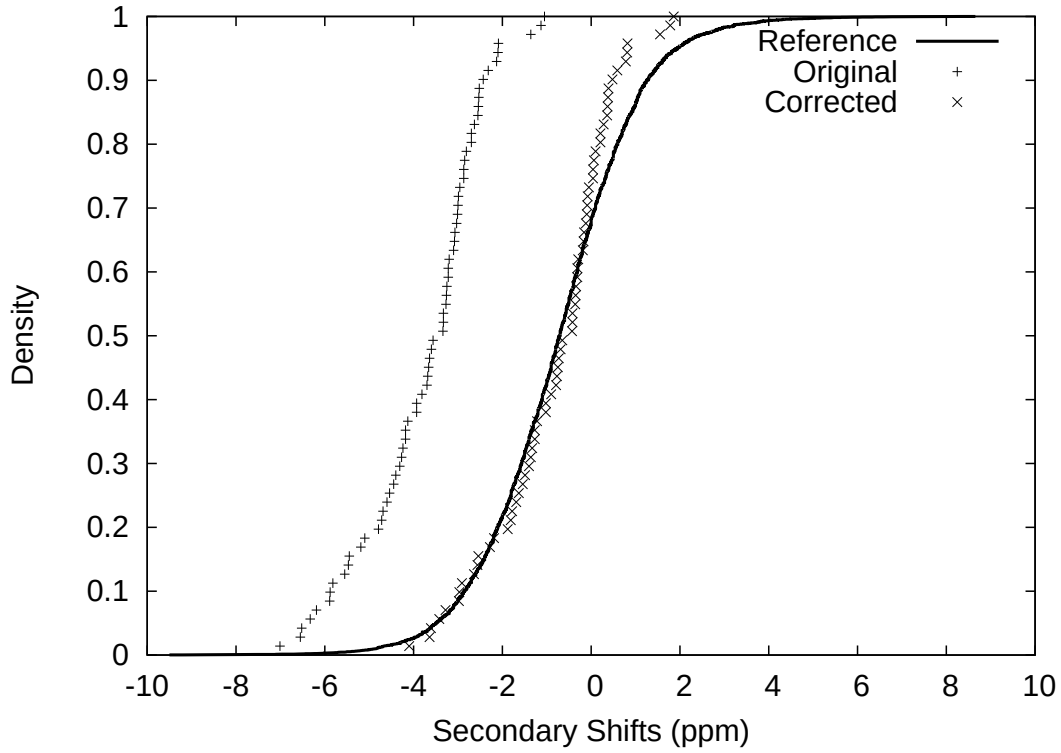


Figure 3.3: Example of the density function of the target's C' shifts for a test protein, $\sigma = (0\%, 37\%, 63\%)$, (original and corrected) and the corresponding reference density function.

Accounting for the secondary structure ratio mentioned above is done by multiplying the density functions for each secondary structure state, derived from the TALOS data set, by the ratio derived from the target protein. The final reference

density function is then calculated as the sum of the three ratio-adjusted density functions.

To evaluate the quality of a certain re-referencing offset, we now calculate the averaged summed distance between the target and the reference density function. This value is inversely proportional to the quality of the proposed offset.

3.2.3 Re-Referencing of Data Sets

The re-referencing is accomplished by searching for the optimal offset over a range defined by the reference distribution, using an increment of 0.1 ppm. Subsequently, all chemical shifts of the data set can be adjusted by adding the determined offset, effectively leading to a data set that is re-referenced to a consistent standard. Additionally, this offset can be used to determine the actual referencing method for a data set.

3.3 Results

The database RefDB [Zhang et al., 2003] uses a structure dependent approach for re-referencing chemical shift data. This is done by comparing a data set to chemical shift data derived from the mapped structure using SHIFTX [Neal et al., 2003]. While this approach is reported to work well and is the established standard, it is limited by the availability of structural data, which is not available for 61% of data sets from the BMRB. Furthermore, some entries in the RefDB still show inconsistencies after re-referencing due to insufficient handling of outliers; chemical shifts that differ from those predicted by SHIFTX by more than four times the expected SHIFTX root mean square error (e.g., 5.0 ppm for $^{13}C_{\alpha}$) do not contribute to the average that is compared to the average of SHIFTX predictions. Data sets which differ strongly from IUPAC standard referencing are therefore re-referenced by an offset that is too small. Here one should keep in mind, that non-standard referencing may not only occurr due to the use of another referencing compound. Offsets may also be induced through calculation errors when the chemical shift data is prepared. The range of these calculation errors may be much larger than the difference between the available reference compounds.

Unlike the RefDB approach, Wang and Wishart [2005] introduced a protocol for adjusting inconsistently referenced chemical shifts that does not depend on structural data. $^{1}H^{\alpha}$-chemical shifts are used to determine the secondary structure of the protein. Subsequently, the re-referencing offset for each chemical shift is derived by comparison to a set of previously published averaged, secondary structure-dependent chemical shifts. These offsets are averaged for each nucleus

over all residues to yield a consensus re-referencing offset for each nucleus. The re-referenced chemical shifts along with the original $^1H^\alpha$-chemical shifts are then used to derive the secondary structure and calculate the re-referencing offset as described before. This last step is iterated twice. CheckShift differs from the approach by Wang and Wishart [2005] in that overall shift distributions are compared rather than individual shifts, and is therefore not exposed to errors in secondary structure prediction for individual amino-acids.

Recently, LACS [Wang et al., 2005b] was developed, a method which calculates re-referencing offsets based on secondary chemical shift values alone. LACS uses linear equations to relate the differences between C_α and C_β shifts to the chemical shift value of C_α, C_β, C' and H_α. By solving these equations, the re-referencing offset for the respective atoms may be calculated. Two constraints have to be fulfilled for LACS to be applicable:

- Chemical shifts for C_α and C_β have to be available.
- C_α and C_β shifts have to be (mis-)referenced in the same way.

In comparison to LACS, CheckShift is not dependent on these constraints, which proves valuable in cases where C_α or C_β shifts are missing or have been referenced differently. Additionally, CheckShift calculates reference corrections for N, which is not possible using the LACS approach.

As it is often hard to check the reliability of chemical shift data, we used a set of 11 target structures (see Table 3.1 for details) provided by the group of Prof. Dr. Horst Kessler from the Technische Universität München for the performance evaluation of CheckShift.

We introduce artificial referencing errors by adding the same offset (artificial error) to all chemical shift values of the targets. This is fair with respect to LACS (as this method is dependent on having the same referencing error for C_α and C_β), however, does not give any advantage to CheckShift as all atoms are processed independently by our method. All multiples of 0.5 in the interval $[-5, 5]$ are used as artificial referencing errors. This way we end up with 220 target chemical shift sets with an artificial error, plus the original 11 chemical shift sets. For each of these chemical shift sets, we calculate the root mean square deviation (RMSD) between the error which was introduced and the negative re-referencing offset calculated by the respective re-referencing methods. The results of this evaluation are shown in Table 3.2. CheckShift strongly outperforms the re-referencing method by Wang and Wishart [2005] and performs equivalently to the LACS approach.

CheckShift's calculations are based on a secondary structure prediction, which is of course not free of error. Therefore, it is interesting to evaluate the dependence of CheckShift on the correctness of the secondary structure assignment. Here, we

Name	Reference	Length	%Helix	%Sheet	%Coil
β-ADT	Heller et al. [2004]	154	27	27	46
HAMP	Hulko et al. [2006]	54	69	0	31
KdpB	Haupt et al. [2006]	136	36	32	32
Mj0056	*EMBO-J, in press*	136	16	41	43
Ph1500N	*unpublished*	83	13	41	46
PhS018	Coles et al. [2006]	92	22	52	26
VatN	Coles et al. [1999]	185	15	36	49
josephin	Nicastro et al. [2005], Mao et al. [2005]	182	38	20	42

Table 3.1: The benchmark set used for performance evaluation. Three unpublished chemical shift sets are not included in the table.

Method	C_α	C_β	C'	N
CheckShift	0.25	0.24	0.55	0.71
Wang and Wishart [2005]	0.81	0.59	1.42	1.12
LACS	0.20	0.20	0.66	n/a

Table 3.2: RMSD of the re-referencing errors.

use 8 target structures from our test set, for which three-dimensional structural information is available (these are the ones listed in Table 3.1). The secondary structure for these targets is calculated using STRIDE. Then a certain percentage of the secondary structure assignments is falsified randomly. Therefore, a certain percentage of the residues in the target are selected randomly. Subsequently, the correct secondary structure assignment for these residues is changed to one of the other two possibilities (e.g., helix might be changed to strand or coil, depending on a random function). This way we generate a set of targets with a secondary structure prediction correctness of 50%, 60%, 70%, 80%, 90%, and 100%. Then, we evaluate CheckShift on all of these sets. The results (shown in Table 3.3) reveal that the quality of C_α and C_β corrections is slightly dependent on the secondary structure assignment. For C' and N offsets hardly any effect may be observed. This proves empirically that CheckShift is stable with repect to errors in secondary structure prediction of up to 50%, as none of the RMSD exceeds 0.7ppm even in the a case of highly unreliable predictions.

✔$_{(H,S,C)}$	C_α	C_β	C'	N
50%	0.53	0.41	0.69	0.48
60%	0.42	0.31	0.53	0.56
70%	0.29	0.26	0.47	0.32
80%	0.33	0.31	0.49	0.36
90%	0.23	0.22	0.44	0.53
100%	0.16	0.21	0.44	0.55

Table 3.3: RMSD of the re-referencing errors for different secondary structure prediction error rates. The first column (✔$_{(H,S,C)}$) shows the percentage of correctly assigned secondary structure.

3.4 Discussion

Correct referencing of chemical shift data is vital for its further use. In the scope of this work, a re-referencing protocol was developed, which does not use structural information, as opposed to established approaches. For this purpose, chemical shifts from a target protein are compared to chemical shift data from a set of correctly referenced proteins by comparing the two datasets' density functions. Subsequently, the target chemical shifts are re-referenced by applying an offset to the chemical shifts of the target. The offset that maximizes the similarity between the target and reference chemical shift data is proposed as the *re-referencing offset*.

By assessing the performance of this approach, it was found the CheckShift performs very well in correcting referencing errors. CheckShift strongly outperforms another structure-independent re-referencing protocol by Wang and Wishart [2005]. The comparison to LACS, a recently proposed re-referencing method, shows that CheckShift performs equivalently. Thereby CheckShift has the advantage of being able to re-reference the chemical shift for each atom independently and to give re-referencing offsets for nitrogen atoms.

3.5 Availability

`http://shifts.bio.ifi.lmu.de/checkshift`

4 SimShift

4.1 Introduction

NMR spectroscopy is one of the most important methods for resolving structures on an atomic level and has been successfully applied to macromolecules such as proteins. Several problems arise on the way from the NMR experiment to the full determination of the 3D coordinates of the structure. One of them is the interpretation of the so-called *chemical shifts*. These are known to inherently carry structural information. It is a difficult task to determine the topology of the protein from the chemical shift data alone. This is therefore usually done by incorporating (human) expert knowledge, in combination with modeling tools and additional experiments — a time consuming process that may take up to several months.

We present an approach (called *SimShift*) to identifying structural similarities among two proteins by searching for similarities in the associated chemical shift sequences. This is done by computing an alignment of the two sequences, the so-called *shift-alignment*. The shift-alignment algorithm will be presented in Sec. 4.3.

The justification of our approach can be seen as follows: The chemical shift for a certain nucleus is influenced by its environment. This is due to the fact that surrounding electron clouds induce a local magnetic field which adds to or subtracts from the field applied in the NMR experiment. This results in different shifts for different environments. In a protein the chemical shift of an atom is influenced by the type of amino acid it is part of *and* the electron clouds of atoms which are close in space due to the tertiary structure of a protein. The influence of the amino acid type may be removed by subtracting the so-called *random coil shifts*, which give an average value for the chemical shift of a specific atom in a certain amino acid. These normalized chemical shift values are called *secondary shifts*. For a short introduction into NMR spectroscopy see Section 2.2.

We will empirically prove the claim that similarity of the shift sequences implies similarity of the respective structures. To do so, in Sec. 4.2 we define a benchmark set which we show to be hard for structure prediction. This set consists of pairs of proteins which have *high* structural similarity (measured by the the MaxSub score [Siew et al., 2000] of their best superposition) but *low* sequence similarity. We choose the MaxSub score as a measure of structural correctness since it is

a good trade-off between the RMSD and the number of aligned residues (see Section 2.11). In analogy to the definition of structural correctness, we define *sequence similarity* as the MaxSub score of the superposition of the residue pairs assigned by an amino acid sequence alignment algorithm. By performing the steps presented in Sec. 4.2, it is possible to generate hard test sets for the evaluation of protein prediction methods in general. Since the scientific community lacks a clearly defined method for deriving benchmark sets for such a task, this can be viewed as another important contribution of our research.

We show in Sec. 4.4 that SimShift is capable of detecting non-trivial structural similarities. SimShift is always better than a mere secondary structure alignment (SSEA), and beats HHsearch (a method that uses both primary and secondary structure information) in more than 50% of all cases. This shows that our alignment quality is situated in the gap between methods that make use of sequence and secondary structure information and high quality structure-structure alignments.

There exists one other prominent approach which aims at inferring structural information at this early stage of an NMR-experiment, namely TALOS [Cornilescu et al., 1999]. The TALOS approach predicts backbone torsion angles from chemical shifts and sequence information by making use of a database of high quality X-ray structures *and* resonance assignments. This works roughly as follows: A sliding window of length three is used to partition the input sequences into triples. For each such a triple the database is then searched for similar triples (in terms of sequence and shifts) for which the torsion angles are known, and the best 10 matches are selected. On basis of their agreement the ϕ and ψ angles are either calculated or the prediction is declined. If any, TALOS gives very accurate predictions, which is the case for about 40% of all residues in the author's benchmark set (or 2/3 after optimization by a human expert). For more detailed information on TALOS see Section 2.4.

SimShift is different from TALOS in the sense that it searches for similar structures rather than *predicting* the structure. It thus enables the construction of a crude model of the query structure, even if there is no "exact" match in the database. This is the reason why we are not bound to using high-resolution X-ray structures as templates, as it is the case for TALOS. Indeed, any structure for which coordinates and chemical shifts are available may be used for comparison. A comparison against TALOS reveals that the SimShift alignments result in significantly better ϕ- and ψ-angle predictions for about half of the targets in our test set.

A typical application of chemical shift alignments may be as follows: Having measured the NMR shift data of a protein with unknown tertiary structure, the obtained sequence is compared to a database of resolved proteins for which shift data is also available. The unresolved protein is likely to be similar in structure

to a protein whose shift values align well. So the three dimensional model of the aligned residues of the known protein is a good starting point for resolving the new structure.

In this chapter we concentrate our research on pairwise chemical shift alignments. Based on the result for pairwise comparisons, we then developed SimShiftDB, a database search tool which fulfills the task sketched in the previous paragraph (see Chapter 5 for details).

4.2 Selection of the Benchmark Set

Several constraints have to be considered for the benchmark set. First, the protein pairs shall be similar in structure, however, amino acid sequence similarity shall be low as otherwise the structural similarity would be easily detectable by an amino acid sequence alignment. Additionally, chemical shift data has to be available for each protein in the benchmark set. The selection of the benchmark set is sketched in Fig. 4.1.

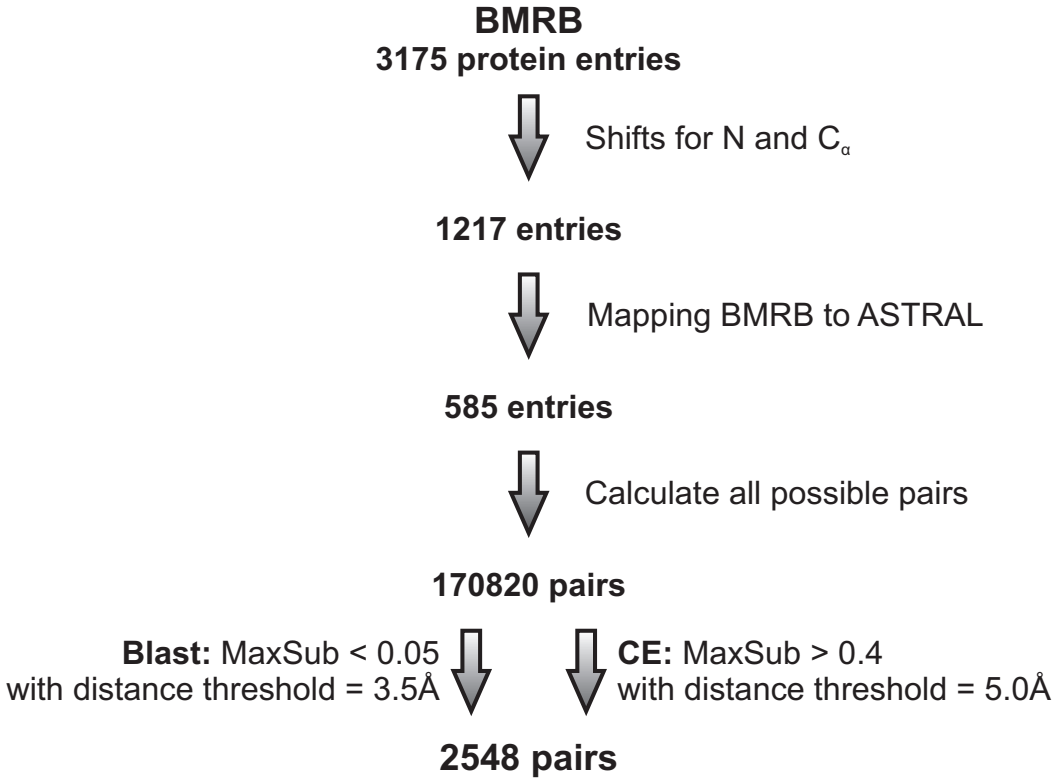


Figure 4.1: Selection of the benchmark set.

4.2.1 Databases Used

All protein entries from BMRB [Seavey et al., 1991] for which N- and C_α-shifts are available are used. The snapshot of the BMRB-database was taken on February 22th, 2005. From 3175 BMRB-entries of the proteins/peptide class, 1217 contained chemical shift values for N and C_α for at least 80 residues.[1] Apart from the information on chemical shifts, each BMRB-entry also contains the corresponding amino acid sequence, though their source (i.e., the protein where the sequence was taken from) is not given in a regular way, sometimes even missing. To identify protein structures corresponding to these amino acid sequences, a BLAST-search [Altschul et al., 1990] against the sequences from the ASTRAL [Chandonia et al., 2004] database is conducted for each BMRB-entry. If the full BMRB sequence can be matched without gaps against an ASTRAL sequence, the corresponding ASTRAL structure is assigned to the BMRB entry. After this selection procedure we ended up with 585 structures that contain chemical shift values for N and C_α and could be mapped to an ASTRAL entry.

As some entries in BMRB match more than one sequence in ASTRAL, one representative structure has to be chosen. This is accomplished by using the AEROSPACI score [Chandonia et al., 2004] provided for each ASTRAL entry. The main contribution to the AEROSPACI score comes from the resolution of the corresponding protein. Higher scores represent better resolutions. Therefore, if more than one sequence from ASTRAL matches one BMRB sequence, the structure with the highest AEROSPACI score is chosen. In general, chemical shift data is not available for every residue in the structure. The matched structures are cut at the beginning and the end to remove overhanging ends, for which no chemical shift data is available.

To calculate the secondary structure assignment (needed in Phase 3 of the algorithm), STRIDE [Frishman and Argos, 1995] was run on all structures in ASTRAL. Via the mapping the assignment was transfered to the corresponding residues in each BMRB entry.

4.2.2 Evaluating the Structural Correctness of Alignments

We will often need to measure the structural correctness of an alignment. The following procedure is always used:

- Extract pairs of aligned residues from the alignment.
- Extract the coordinates of the C_α atoms from the tertiary structures.

[1] This threshold was chosen to exclude BMRB-entries consisting of just one secondary structure element.

- Superimpose the two point sets. Here we use the superposition algorithm by Kabsch [1978] as it calculates the optimal superposition as according to the RMSD between the aligned residues.
- Calculate the MaxSub-score [Siew et al., 2000] using a certain distance threshold.

The latter is used as a measure of structural correctness of the alignment. This procedure is chosen as in our work we are interested in detecting *structural similarity*. Therefore, it is important to evaluate the *structural* correctness of the alignments created. As we first optimize the superposition of the corresponding protein structures based on RMSD (using the algorithm by Kabsch [1978]), but evaluate the quality of the alignment using the MaxSub score, we additionally avoid training SimShift to work well with just one specific scoring system.

4.2.3 Defining a Benchmark Set

Our aim is to show the algorithm's ability to identify structural similarities in pairs of proteins where sequence similarity is low. To create a test set with these properties, we first compute all possible combinations of our 585 structures from Sect. 4.2.1. Now, we select pairs that fulfill the following constraints:

Low predictability from the sequence. We calculate a BLAST pairwise alignment for all pairs. If a BLAST alignment has been found, we evaluate the structural correctness of the alignment with the method from Sec. 4.2.2. We keep all pairs that either do not have a BLAST alignment or whose BLAST alignment has a MaxSub-score ≤ 0.05, where the distance threshold is 3.5Å.

Existence of structural similarity. The pairs to be finally used should have some detectable structural similarity, despite of their low sequence similarity. To identify such proteins pairs we calculate CE-alignments for all remaining test pairs with the method presented by Shindyalov and Bourne [1998]. The correctness of the alignment is again evaluated as described in Sec. 4.2.2. We keep those pairs with a MaxSub-score > 0.4, where the distance threshold is 5.0Å.

MaxSub is insensitive to small variation in the distance cutoff (see Section 2.11). The value of 3.5 is chosen according to the recommendation by Siew et al. [2000]. For the second constraint, the cutoff value is slightly relaxed to include also more distant structural similarities. The MaxSub cutoffs are adjusted to achieve a reasonable size of the benchmark set, thereby being sure that a MaxSub score of ≤ 0.05 does definitely describe a completely insignificant alignment, and that

a MaxSub score of > 0.4 means there is definitely some detectable structural similarity. The set of pairs that passed these two criteria consists of 2548 pairs which are built from 417 structures. Their average number of residues is 117, with a standard deviation of 38. As we imposed a minimum length restriction (see Section 4.2.1) none of the structures has less than 80 residues.

4.3 The Shift-Alignment Algorithm

For the algorithm presented in this section we will use the shift-values of the C_α- and the N-atoms (from the backbone of the protein). The algorithm takes two amino-acid sequences $s = s_1 \dots s_n$ and $t = t_1 \dots t_m$ and the shift-values of the respective C_α- and the N-atoms and returns a list of aligned amino-acids. This is done in three phases that are explained in the next subsections.

4.3.1 Phase 1: Calculation of the Shift-Difference Matrix

For the shift-values of the C_α-atoms, we compute the distance for each possible pairing of the amino-acids and store the result in a shift-difference matrix M_{C_α}, M_N, and i.e., for all $1 \leq i \leq m$ and $1 \leq j \leq n$, we calculate

$$\begin{aligned} M_{C_\alpha}[i][j] &= |\overline{\delta}_{C_\alpha}(t_i) - \overline{\delta}_{C_\alpha}(s_j)|, \\ M_N[i][j] &= |\overline{\delta}_N(t_i) - \overline{\delta}_N(s_j)|, \end{aligned} \tag{4.1}$$

where $\overline{\delta}_{C_\alpha}(t_i)$ and $\overline{\delta}_N(t_i)$ are the secondary shifts of the C_α and N atoms of residue i in sequence t, respectively. The shift-difference matrix M_N is computed accordingly for the $\overline{\delta}_N$-values. We stress the fact that the shift-difference matrices are only of conceptual nature and need not be calculated explicitly. Nevertheless, they facilitate the understanding of the algorithm. The secondary shift values are obtained using the random coil shifts from Wishart et al. [1995b]. For the ^{15}N random coil shift in Proline (not available from Wishart et al. [1995b]), we take the results from Braun et al. [1994].

4.3.2 Phase 2: Find Good Blocks

We now wish to find a set of blocks $\{b_h\}$ that represents "good" local alignments (without gaps) of substrings from s and t. More formally, a block b is defined as a triple (i, j, k) with $1 \leq i \leq m - k + 1$ and $1 \leq j \leq n - k + 1$, where the intended meaning is that $t_i \dots t_{i+k-1}$ aligns with $s_j \dots s_{j+k-1}$. For simplicity of notation, for a given block $b = (i, j, k)$, we define the *block extents* $X_{\min}(b) = j$, $Y_{\min}(b) = i$, $X_{\max}(b) = j + k - 1$, and $Y_{\max}(b) = i + k - 1$.

Two restrictions are placed on these blocks. First, they should fulfill a minimum length criterion, so we require that $k \geq l$ for some minimum length l. Second, all aligned amino-acids should have "similar" shift-values for both the C_α- and the N-atom, i.e. $M_{C_\alpha}[i+p][j+p] \leq \gamma_{C_\alpha}$ and $M_N[i+p][j+p] \leq \gamma_N$ for all $0 \leq p < k$. We further require the extent of blocks to be maximal, i.e. $M_{C_\alpha}[Y_{\min}(b)-1][X_{\min}(b)-1] > \gamma_{C_\alpha}$ or $M_N[Y_{\min}(b)-1][X_{\min}(b)-1] > \gamma_N$ and likewise for the other end of the block $(X_{\max}(b)+1, Y_{\max}(b)+1)$. The values γ_{C_α} and γ_N are called *cutoff*-parameters. A graphical depiction of this concept can be found in Fig. 4.2. For the rest of this section, we denote by n' the number of blocks that have been found in Phase 2.

4.3.3 Phase 3: Concatenation of Blocks

In this step, multiple local alignments are concatenated to a global alignment consisting of more than one block. To find the best global alignment, a positive score (representing the block's global correctness) is first associated with each block. To calculate this score, we re-use the idea of secondary shifts; here, however, we normalize not only according to the amino acid type, but also according to the protein's secondary structure. This increases the influence of long-range interactions on the score. For normalization, we calculate z-scores using the averaged β-strand, random-coil, and α-helix shifts, as well as the according standard deviations from Wang and Jardetzky [2002]. To calculate a specific block score, we calculate the N and C_α differences between the normalized chemical shifts and sum them over all residue pairs in the block. A more formal definition of the block score is given in the following paragraph.

Defining ζ to be the chemical shift value that is normalized according to secondary structure *and* amino acid type, we set

$$M = \max_{\substack{c(i,j,k)\in\{b_1,\ldots,b_{n'}\},\\ r\in\{0,\ldots,k-1\}}} \left\{ |\zeta_N(t_{i+r}) - \zeta_N(s_{j+r})| \; + |\zeta_{C_\alpha}(t_{i+r}) - \zeta_{C_\alpha}(s_{j+r})| \right\},$$

which is the maximum of all pairwise differences in the set of blocks that has been found in Phase 2. We then define the score s of a block (i, j, k) as

$$s(i,j,k) = \sum_{r=0}^{k-1} \left(M - |\zeta_N(t_{i+r}) - \zeta_N(s_{j+r})| - |\zeta_{C_\alpha}(t_{i+r}) - \zeta_{C_\alpha}(s_{j+r})| \right).$$

The effect of this formula is that the pair scores are inverted and moved to the positive range to associate the highest score with the best pair. Using these scores, we apply the following algorithm to identify the highest scoring block chain.

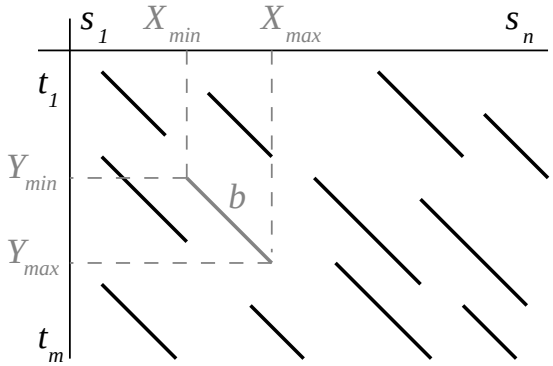


Figure 4.2: Blocks in the alignment matrix: thick lines represent "good" local alignments in the sense of Sec. 4.3.2. The block extents for a block b are also depicted.

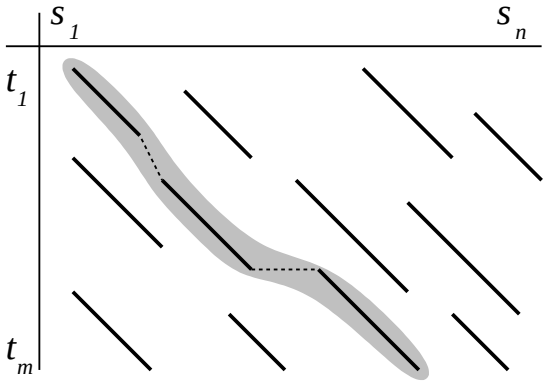


Figure 4.3: A global alignment (highlighted) constructed from good blocks. Dotted lines represent gaps in either of the sequences.

For a given set of blocks $\{b_h\}$, a *block chain* B is defined as a non-overlapping sequence of blocks $b_{h_1}, \ldots, b_{h_r}$, where "non-overlapping" means that $X_{\max}(b_{h_q}) < X_{\min}(b_{h_{q+1}})$ and $Y_{\max}(b_{h_q}) < Y_{\min}(b_{h_{q+1}})$ for all $1 \leq q \leq r$. The block score is extended in a natural way to block chains by setting $s(B) = \sum_{q=1}^{r} s(b_q)$. The *optimal* block chain is the one with maximal score. See Fig. 4.3 for an example. Further, the *block extents* are extended to block chains by defining $Y_{\min}(B)$ to be the minimum of all $Y_{\min}$'s in B, and likewise for the other extents $Y_{\max}, X_{\min}$ and $X_{\max}$.

Algorithm 1 is used to compute the optimal block chain for a set of n' blocks. It is a straightforward adaption of the algorithm by Joseph et al. [1992].[2] We note that with an efficient implementation of the set D the running time of Algorithm 1

[2]In line 11 of Alg. 1 we corrected a slight mistake in the original algorithm Joseph et al. [1992] by comparing $Y_{\min}(B_c)$ to $Y_{\mathbf{min}}(b_j)$ instead of $Y_{\mathbf{max}}(b_j)$.

is $O(n' \log n')$. As phases 1 and 2 can both be implemented in $O(nm)$, the total running time of the shift-alignment algorithm is $O(nm + n' \log n')$. Although this term may be as high as $O(nm \log (nm))$, it will be $O(nm)$ even for slightly "restrictive" choices of the parameters l, γ_{C_α} and γ_N, because with few blocks calculated in Phase 2 the $n' \log n'$-term is asymptotically less than the nm-term.

Algorithm 1 Chaining Algorithm

```
1: initialize D as an empty list
2: let X = {x_1, ..., x_2n'} be the list of points X_min and X_max, sorted in decreasing
   order
3: for i = 1, ..., 2n' do
4:    if x_i = X_max(b_j) for some j then
5:       B_c ← first block-chain in D such that Y_min(B_c) > Y_max(b_j)
6:       if B_c = ∅ then
7:          B_j = {b_j} {start new block chain}
8:       else
9:          B_j = {b_j} + B_c
10:   else if x_i = X_min(b_j) for some j then
11:      B_c ← first block-chain in D such that Y_min(B_c) ≥ Y_min(b_j)
12:      if B_c = ∅ or s(B_c) ≤ s(B_j) then
13:         insert B_j into D s.th. ∀ 1 ≤ k < |D|:
            Y_min(D[k]) ≤ Y_min(D[k + 1]) and
            s(D[k]) ≥ s(D[k + 1])
14:         for all B_↓ ∈ D\B_j do
15:            if Y_min(B_↓) ≤ Y_min(B_j) and
               s(B_↓) < s(B_j) then
16:               remove B_↓ from D
17: return D
```

4.3.4 Parameter Optimization

The performance of the algorithm is highly dependent on the choice of the minimum length restriction l and the cut-off parameters γ_{C_α} and γ_N. Since it is hard to predict which combination of the three values yields the best alignments, we try all combinations of the three parameters in the range $l \in \{3, 4, \ldots, 17\}$ and $\gamma_{C_\alpha}, \gamma_N \in \{2.0, 2.5, 3.0, \ldots, 10.0\}$. The average difference between the MaxSub-scores of SimShift and CE is used for ranking the parameter combinations.

Several combinations of the parameters yield equally good scores. Among those, we inspected some by hand and found a clear influence of the parameters on the RMSD and the length of the alignment: Lowering γ_{C_α} and γ_N and increasing l yields shorter alignments with a better RMSD, whereas raising the cut-offs

and lowering the minimal block length yields longer alignments with a higher RMSD. In the following, we use the cut-off values $\gamma_N = 5.5$, $\gamma_{C_\alpha} = 3.5$ and a minimum blocks length of 12, which seems to be a reasonable tradeoff between specificity and sensitivity. In practice, one might start searching a database of secondary shift sequences with a target sequence using very strict parameters. If no satisfactory answer is found, one might start loosening parameters. That way, some similarity to another structure may be found in any case; however, one has to keep in mind that the probability of achieving a correct structural alignment declines.

4.4 Results

4.4.1 Comparison to SSEA and HHsearch

We compare SimShift to SSEA [Fontana et al., 2005] and HHsearch [Söding, 2005]. The former is used to rule out the possibility that SimShift does a mere alignment of secondary structure elements, the latter because it is a state-of-the-art method that incorporates both sequence and secondary structure information. We use the benchmark set from Sect. 4.2 for comparison.[3] In each pair of this set, one structure is used as a template, the other as a target. For SSEA and HHsearch, the secondary structure of the target is computed by PSIPRED [Jones, 1999], whereas for SimShift it is computed by the method of [Wishart et al., 1992]. In a second comparison, we used the predictions given by [Wishart et al., 1992] for all methods, however, this yielded worse results for both SSEA and HHsearch. The secondary structure of the template is calculated by STRIDE for all methods. We only compare alignments where all three methods produce a non-empty result. Therefore, the number of pairs is reduced to 1373. For the comparison a MaxSub score with a distance threshold of 5.0Å is used.

Fig. 4.4 and Fig. 4.6 show the percentage of pairs where SimShift is better than SSEA (top line) or HHsearch (bottom line). SimShift is substantially better than SSEA, revealing that the method achieves more than a mere secondary structure element alignment. Regarding the comparison to HHsearch, note that in the region where the pairs do not have a global structural similarity ($.4 \leq$ CE MaxSub $< .58$), SimShift is significantly better. However, with rising structural similarity, HHsearch plays off its strength.

We are further interested in the gain in alignment quality, one can achieve by analyzing chemical shifts in addition to primary and/or secondary structure. Therefore, we investigate those cases where SimShift is better than both of the

[3] This is admissible because the cutoff parameters for SimShift were optimized against the CE-alignment (see Sec. 4.3.4), whereas here we compare to different methods.

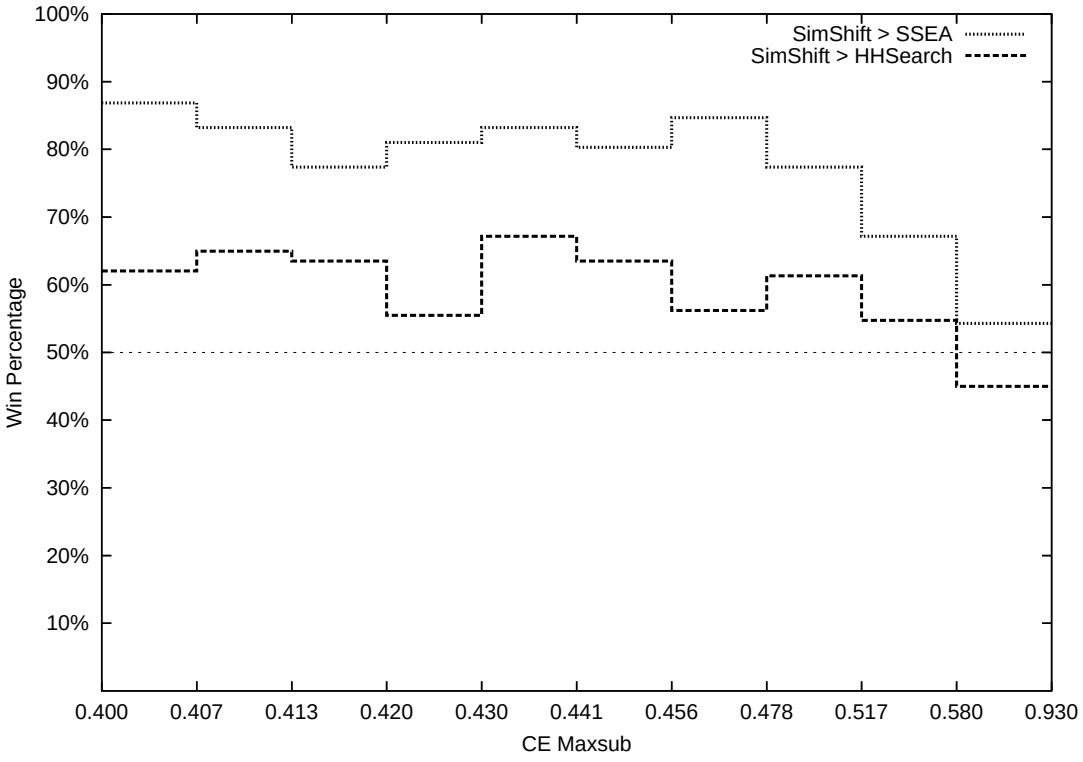


Figure 4.4: The performance of SimShift compared to SSEA and HHsearch. The secondary structure for SSEA and HHsearch is calculated using PSIPRED. The actual structural similarity of the pairs is plotted against the percentage of alignments where SimShift achieves a higher MaxSub score than SSEA or HHsearch, respectively. Note that each segment is the average over 137 pairs, so the step-width of the x-axis is not linear.

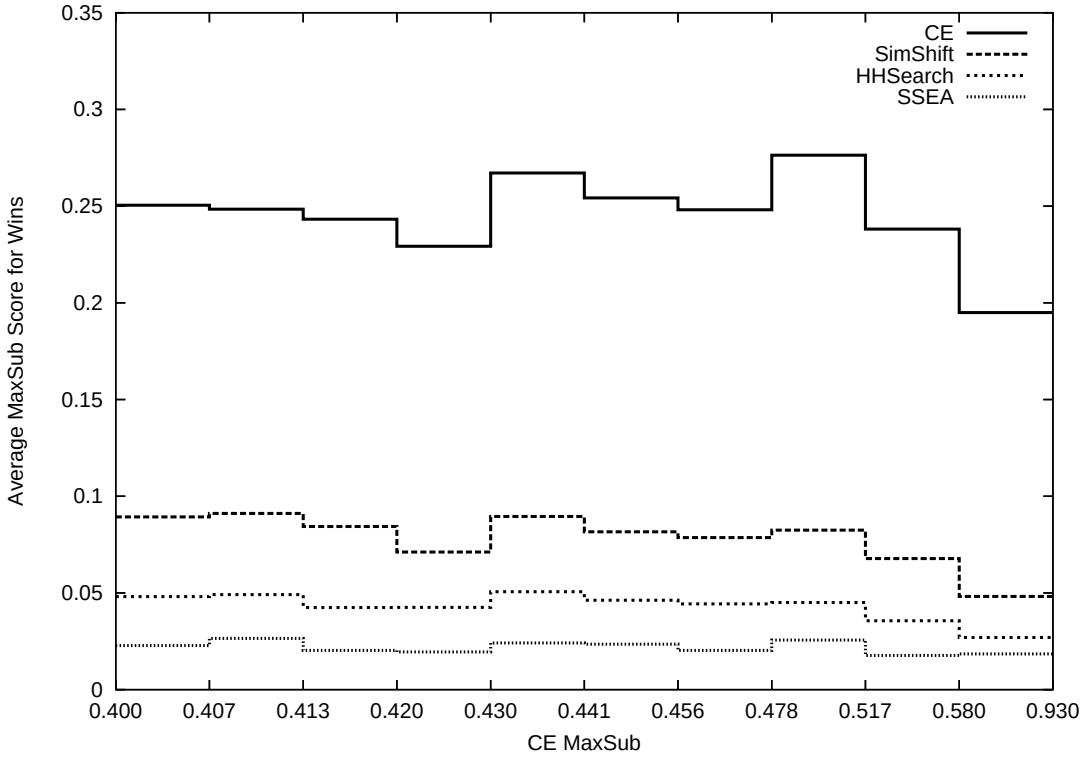


Figure 4.5: Average MaxSub scores for CE, SimShift, HHsearch and SSEA for pairs where SimShift outperforms the other two methods. The secondary structure for SSEA and HHsearch is calculated using PSIPRED. The x-axis is equivalent to Figure 4.4.

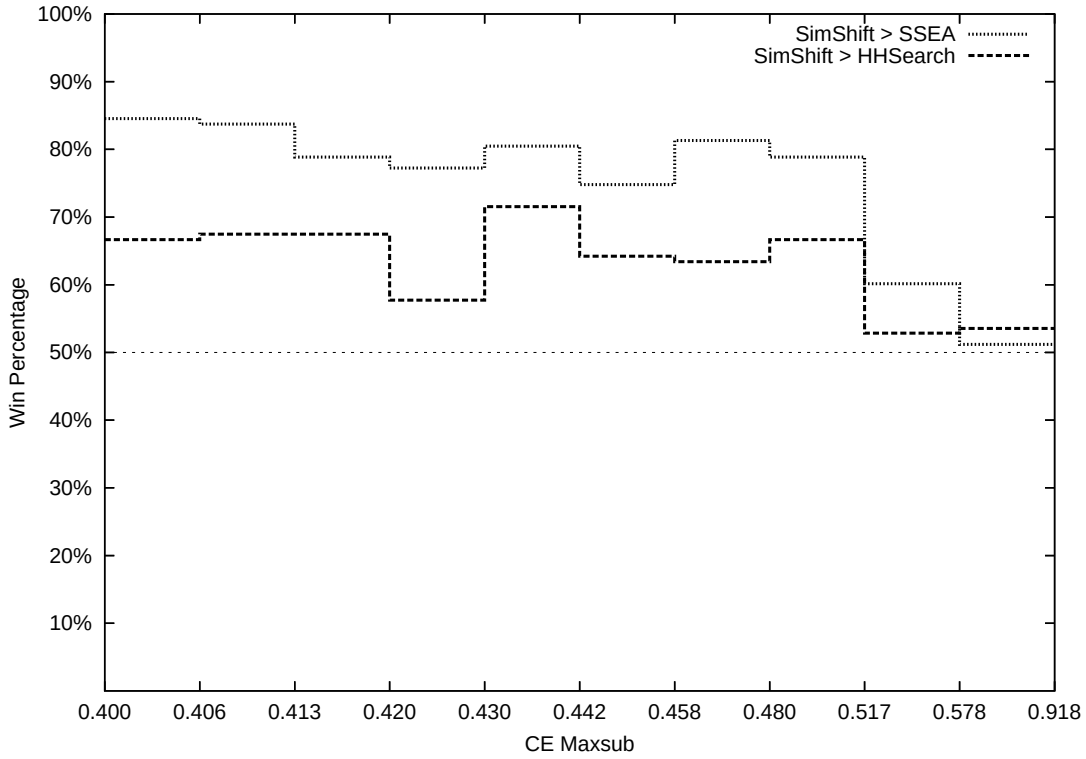


Figure 4.6: The performance of SimShift compared to SSEA and HHsearch. The secondary structure for all methods is calculated using CSI. The actual structural similarity of the pairs is plotted against the percentage of alignments where SimShift achieves a higher MaxSub score than SSEA or HHsearch, respectively. Note that each segment is the average over 123 pairs, so the step-width of the x-axis is not linear.

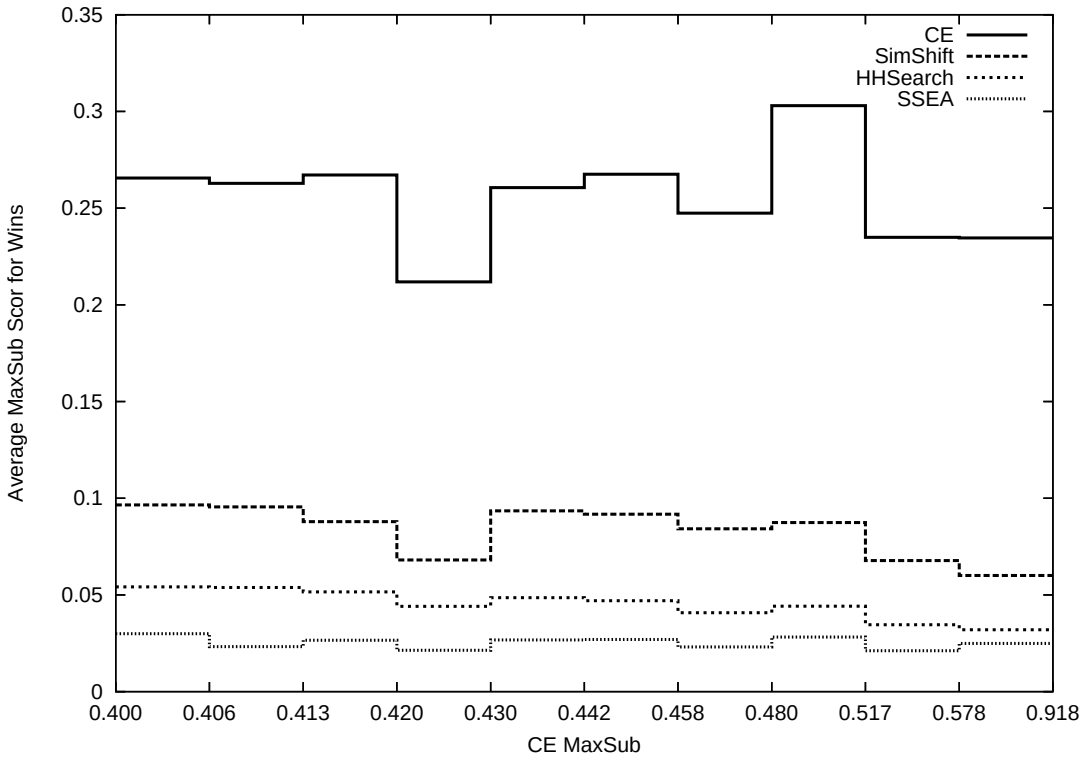


Figure 4.7: Average MaxSub scores for CE, SimShift, HHsearch and SSEA for pairs where SimShift outperforms the other two methods. The secondary structure for all methods is calculated using CSI. The x-axis is equivalent to Figure 4.6.

other methods. The results can be seen in Fig. 4.5 and Fig. 4.7 (bottom 3 lines). It shows that the average MaxSub score of the alignments made by our method is much better than the average scores of SSEA's and HHsearch's alignments. In fact, SimShift is about 3 times better than SSEA, and about twice as good as HHsearch.

We further plot the average MaxSub scores of the respective CE-alignments in the graph of Fig. 8 (top line). It is interesting to see that it is much lower than the average score of the whole segments on the x-axis. This emphasizes the fact that SimShift is especially useful in cases where structural similarity is not very high and the performance of other methods decreases.

4.4.2 Comparison to TALOS

As TALOS predicts backbone angles rather than producing an alignment, it was impossible to include it in the tests of the previous section. Nevertheless, as the ultimate goal of this work is to be able to do homology modeling based on chemical shift alignments, it is important to evaluate if there are cases where SimShift is able to provide higher quality torsion angle predictions than TALOS.

Therefore, we split the proteins in the benchmark set into two parts: Of each pair, one protein is classified as being the modeling target and the other is used as a potential template. This leaves us with a set of 363 target proteins, each one having at least one associated template structure.

Because the true structure of our targets is known, we can compare the RMSDs of the backbone angles of the best alignment produced by SimShift to the predictions made by TALOS. Thereby, we use only residues where both methods provide torsion angles. Of all 363 targets, 178 have a better RMSD for both the ϕ- and the ψ-angle. The average RMSD-difference for those where SimShift is better is 18.18° for ϕ, and even 45.58° for ψ. This shows that SimShift can be useful for assisting the structure resolution process even in the presence of TALOS.

4.5 Discussion

We aimed at answering the question: "Is it possible to create structurally correct alignments from chemical shift data alone, when sequence similarity is low?" We argued that this is indeed the case. Through the comparison to other methods we also motivated that information about long range interactions can be extracted from chemical shift data and may be used to create structurally meaningful alignments.

The shift data used here is derived from the BMRB, which is known to contain high quality as well as low quality entries. We were interested in the performance of our approach on experimental data, we therefore did not include any intermediate processing steps. Additionally, because there is only a limited number of proteins with associated chemical shifts, it is *not* advisable to reduce this set even more by restricting oneself to confirmed high quality entries. As the performance presented here was achieved on shifts probably containing erroneous data, one can expect even more accurate alignments when using curated shift data.

What has been presented is a first step towards automating the structure determination process with NMR spectroscopy. Chemical shift alignments can be a useful tool for the spectroscopist who starts searching a database of chemical shifts before performing additional experiments. If similarities can be identified a model for the protein of interest may be created. Through comparison to NOE maps, for example, it is possible to validate (or invalidate) the model.

There is still some work to do towards automating structure determination. In the following chapter we present SimShiftDB, a database search tool based on chemical shift alignments. Using the similarities identified by our database search we are able to infer structural information from database proteins to the target protein we are working on. We also apply a statistical model to assess the significance of each similarity identified.

5 SimShiftDB

5.1 Introduction

NMR Spectroscopy is an established method to resolve protein structures on an atomic level. The NMR structure determination process consists of several steps, varying in complexity. A quantity that is measured routinely in the beginning is the chemical shift. Chemical shifts are available on a per atom basis and inherently carry structural information. Chemical shifts in general do not suffice to calculate the structure of biological macromolecules, such as proteins. Additional experiments of increased complexity and human expert knowledge are necessary to obtain the solution.

In Chapter 4, the performance of a pairwise chemical shift alignment algorithm was evaluated. We were able to show that it is indeed possible to utilize the information hidden in the chemical shift data for constructing structurally meaningful alignments. Now we present a method (called *SimShiftDB*) that searches for similarities between a target protein with assigned chemical shifts and a database of template proteins for which both chemical shift data and 3D coordinates are available. The alignment algorithm used in the previous chapter was adapted to fit the requirements of database searching. Also additional constraints derived from the template structure have been incorporated into the calculations.

For each target-template-pair we calculate a chemical shift alignment. These alignments map a set of residues from the target to a set of residues from the template structure. Therefore, we can build a structural model for the aligned residues from the target based on the coordinates of the associated residues from the template. To give the spectroscopist the possibility to judge over the statistical significance of a certain alignment with shift similarity score S, we calculate the expectation of the number of alignments with score $\geq S$ occurring by chance.

To evaluate the performance of our approach, we compare the backbone angle prediction quality of our method to 123D [Alexandrov et al., 1996], a threading approach, and to TALOS [Cornilescu et al., 1999], which tries to calculate backbone angles from the amino acid sequence and associated chemical shift data. We are able to prove empirically that 123D is outperformed significantly by our method. When comparing to TALOS, SimShiftDB performs significantly better for 36% of the target's residues. Our result suggests that both TALOS and

SimShiftDB have their strengths and, therefore, should be used in parallel in the NMR structure determination process.

In the following we describe the template database, the chemical shifts substitution matrices used for scoring chemical shift alignments, the calculation of the expected value and the SimShiftDB algorithm. Afterwards, the results on our test set will be presented and discussed.

5.2 The Template Database

The BMRB [Seavey et al., 1991] is the main repository for protein chemical shift data. However, to date there are only 3750 structure in the *proteins/peptide* class, which corresponds to about 10% of the structures deposited in PDB. Additionally, there is no standard set of chemical shifts which has to be available for each entry, e.g., one entry may contain only ^{1}H chemical shifts while for a different one just ^{15}N chemical shifts are available. Finally, as pointed out by Zhang et al. [2003] various errors occur in the data. These problems led us to the conclusion to use a different template database.

Chemical shifts for all structures in the ASTRAL [Chandonia et al., 2004] database are calculated using SHIFTX [Neal et al., 2003]. Chemical shifts predicted with SHIFTX correlate strongly with measured data. It is also shown that the agreement between observed and calculated chemical shifts is an extremely sensitive measure to assess the quality of protein structures (for more information on SHIFTX see Section 2.3). This approach leaves us with a database containing 64,839 proteins with known 3D structure and associated chemical shifts.

5.3 Substitution Matrices for Shift Data

In order to identify pairs of amino acids with associated chemical shifts which are likely to be structurally equivalent, we derive substitution matrices for chemical shifts using the modus operandi described by Henikoff and Henikoff [1992]. Therefore, a *Standard of Truth* is needed for the calculation of the matrix values. We rely on the DALI database [Holm and Sander, 1996] containing a set of 188,620 structure-based alignments. Through a sequence-similarity-search [Altschul et al., 1990] we map all sequences being part of a DALI alignment to our template database.

For each amino acid, we calculate the minimal and maximal chemical shift of $^{1}H_{\alpha}$, $^{13}C_{\alpha}$, $^{13}C_{\beta}$, and ^{13}C in our template database. Then we divide the range between minimum and maximum into two equal parts. This enables us to classify each chemical shift as either *weak* (situated in the first part of the range) or *strong*.

It is convenient to define a new alphabet on proteins with associated chemical shift sequences, namely Σ^S. A letter $\mathcal{A}$ in this alphabet is a tuple (a, s_1, s_2, s_3, s_4), where a is the corresponding amino acid identifier and s_1, s_2, s_3, s_4 are the classifications for the corresponding shifts for $^1\mathrm{H}_\alpha$, $^{13}\mathrm{C}_\alpha$, $^{13}\mathrm{C}_\beta$, and $^{13}\mathrm{C}$, respectively.

We derive the relative frequencies of each of the letters in the template database, denoted by $p_\mathcal{A}$. Additionally, we calculate the relative frequencies of all substitution events, denoted by $q_{\mathcal{A},\mathcal{B}}$, which is the relative frequency of letters $\mathcal{A}$ and $\mathcal{B}$ being aligned in the DALI database. To account for the bias of overrepresented folds, we use pseudo counts to give each fold type an equal weight. To do so, each alignment is identified with the fold type associated to the first sequence according to the SCOP [Murzin et al., 1995] classification.

Then, we calculate the well-known log-odds scores [Henikoff and Henikoff, 1992]

$$o_{\mathcal{A},\mathcal{B}} = \log\left(\frac{q_{\mathcal{A},\mathcal{B}}}{e_{\mathcal{A},\mathcal{B}}}\right), \tag{5.1}$$

where

$$e_{\mathcal{A},\mathcal{B}} = \begin{cases} 2 * p_\mathcal{A} * p_\mathcal{B} & \text{if} \quad \mathcal{A} \neq \mathcal{B}, \\ p_\mathcal{A}^2 & \text{otherwise.} \end{cases} \tag{5.2}$$

Finally the log-odds scores are multiplied with a normalization factor η and rounded to the nearest integer. The shift substitution matrix entries $s_{\mathcal{A},\mathcal{B}}$ are then formally defined as

$$s_{\mathcal{A},\mathcal{B}} = \lfloor o_{\mathcal{A},\mathcal{B}} * \eta + 0.5 \rfloor. \tag{5.3}$$

Here the parameter η is set to 10. This value was chosen based on a thorough inspection of the values $o_{\mathcal{A},\mathcal{B}}$, thereby trying to sacrifice as little information as possible.

5.4 E-Values for Chemical Shift Alignments

A shift alignment produced by SimShiftDB is a set of local ungapped alignments which do not overlap. Karlin and Altschul [1993] derive a p-value for multiple ungapped alignments which may be ordered consistently (see page 5874, section *Consistently Ordered Segment Pairs in Sequence Alignments* for details). For this p-value two statistical parameters (λ, κ) have to be calculated.

We use the method described by Karlin and Altschul [1990] to obtain these parameters. Formally λ is defined as the unique positive solution of the equation

$$\sum_{\mathcal{A},\mathcal{B} \in \Sigma^S} p_\mathcal{A}^T * p_\mathcal{B}^D * e^{\lambda * s_{\mathcal{A},\mathcal{B}}} = 1, \tag{5.4}$$

where $p^T_{\mathcal{A}}$ is the probability that the letter $\mathcal{A}$ occurs in the target sequence and $p^D_{\mathcal{B}}$ is the probability that letter $\mathcal{B}$ occurs in the template database.

The parameter κ is calculated as

$$\kappa = e^{\gamma} * \frac{\delta}{(1 - e^{-\lambda * \delta}) * E[S[1]e^{\lambda * S[1]}]}, \tag{5.5}$$

with

$$\gamma = -2 * \sum_{k=1}^{\infty} \frac{1}{k} (E[e^{\lambda * S[k]} | S[k] < 0] + P(S[k] \geq 0)) \tag{5.6}$$

and

$$\delta = \gcd\{s_{\mathcal{A},\mathcal{B}} \mid \mathcal{A}, \mathcal{B} \in \Sigma^S\}. \tag{5.7}$$

Here, $S[k]$ is a random variable representing the sum of the pair scores of an alignment of length k. For further details, we refer to [Karlin and Altschul, 1990].

Using λ, κ as described above, we can normalize the score of an ungapped chemical shift alignment A as follows. Let S_A be the sum of the pairwise scores of the aligned letters from Σ^S. The normalized score S'_A is then defined as

$$S'_A = \lambda * S_A - \ln(n * m * \kappa), \tag{5.8}$$

where n is the length of the target sequence and m is the length of the template sequence.

According to Karlin and Altschul [1993], we can calculate the probability that a number of consistently ordered alignments $A_1, \ldots, A_r$ with summed normalized score at least T' occurs by chance as

$$P(T', r) = \int_{T' + \ln(r!)}^{\infty} \frac{e^{-t}}{r!(r-2)!} \int_0^{\infty} y^{r-1} e^{-e^{\frac{y-t}{r}}} dy\, dt, \tag{5.9}$$

with

$$T' = \sum_{i=1}^{r} S'_{A_i}. \tag{5.10}$$

Note that we start integrating not from T', but we add the value $\ln(r!)$. This is due to the fact that in our case the ungapped alignments are ordered consistently. The original theory is based on the assumption that the r ungapped alignments need not be ordered. As we apply the additional constraint of consistent ordering, we effectively divide the solution space by $r!$. This is accomplished by shifting the lower bound of the integral by $\ln(r!)$, as due to the properties of the p.d.f. this divides $P(T', r)$ by $r!$.

However, one problem remains. $P(T', r)$ does not take into account the database size. Therefore, we additionally calculate the expected number of alignments in our search space with a score not less than the score of the alignment of interest:

$$E(T', r) = P(T', r) * \frac{N}{m}. \tag{5.11}$$

Here m is the length of the template sequence and N is the number of letters in the database. $E(T', r)$ is the e-value we use to assess the statistical significance of the chemical shift alignments.

5.5 The Shift Alignment Algorithm

We design a two step algorithm to build a chemical shift alignment for two sequences from the alphabet Σ^S. Initially, a set of local ungapped alignments is constructed. Then we search for a best legal combination of a subset of these alignments.

5.5.1 Step 1: Calculate local alignments

We construct the pair score matrix containing scores for all pairs of letters $\mathcal{A}$ and $\mathcal{B}$, where $\mathcal{A}$ is a letter of the target and $\mathcal{B}$ is a letter of the template sequence. The score for each pairing is $s_{\mathcal{A},\mathcal{B}}$ as defined in Equ. (5.3). Then we apply an algorithm [Ruzzo and Tompa, 1999] which identifies all maximal scoring subsequences (*MSS*) on each diagonal in linear time. An MSS is defined as follows.

Definition 2. *Let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $a' = (a_i, \ldots, a_j)$ be a subsequence of a. a' is a maximal scoring subsequence if and only if*

(1) All proper subsequences of a' have lower score.

(2) No proper supersequence of a' contained in a satisfies (1) .

These conditions uniquely define all MSS in a given sequence. Additionally, it is easily proved that MSS may not overlap.

Here we choose MSS instead of the cutoff approach used in Chapter 4 as with the MSS it is possible to bridge few residues with low pair score, as long as the summed score of the diagonal segment is maximal. This strongly increases the sensitivity of the first step of the algorithm An MSS can also be interpreted as a local ungapped alignment and will be called a *block* from now on. Note that the algorithm only identifies MSSs with score greater than zero. Additionally, we remove all MSSs of length ≤ 5. This cutoff is chosen as similar regions of

length ≤ 5 are likely to occur very often, thereby having a low significance. In the following chapter it is actually shown that SimShiftDB is stable with respect to small variations in the minimum length restriction.

5.5.2 Step 2: Identify the best legal combination

We now build a DAG (directed acyclic graph) in which the blocks correspond to the nodes in the graph. Two blocks may be combined if they are connected by an edge in this graph. Two constraints have to be fulfilled for two blocks (B_1 and B_2) to be connected by an edge from B_1 to B_2:

1. B_1 and B_2 may **not** overlap, neither in the target nor in the template sequence. Additionally, B_1 has to appear before B_2 in the target as well as in the template.

2. Let **d** be the number of residues in the target sequence between the end of B_1 and the beginning of B_2. Let **L** be the last residue from the first block in the template sequence and **F** be the first residue from the second block in the template sequence (see Fig. 5.1). We require the residues **L** and **F** not to be further apart in the structure than the maximal distance that could be bridged by a polypeptide chain of **d** residues. Here it is assumed that the maximal distance between two C_α atoms in the polypeptide chain is 4.0Å.

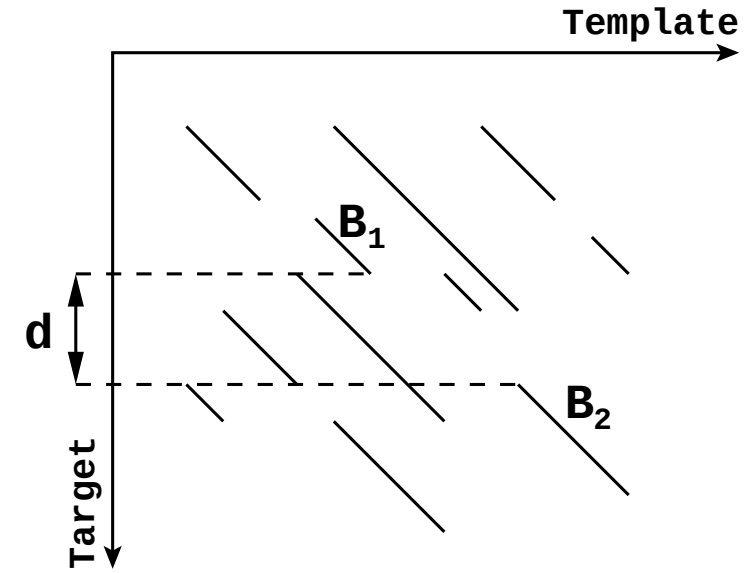
We also add an artificial start node to the DAG from which every other node may be reached.

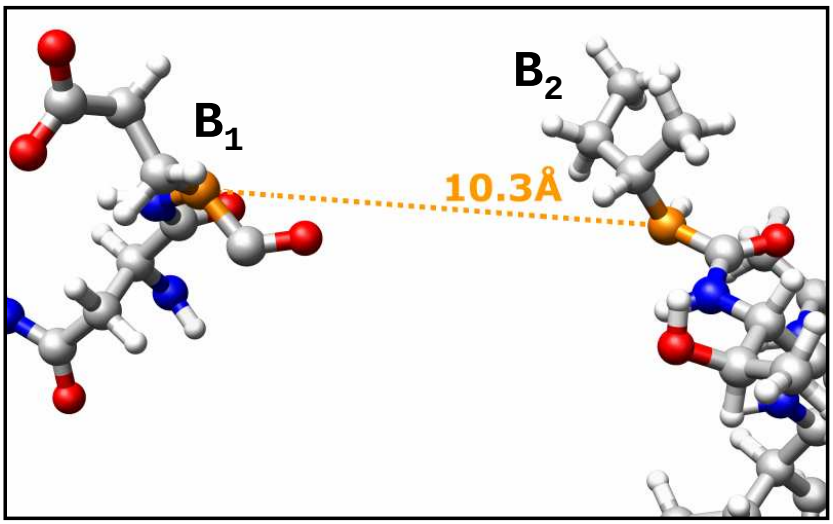
Fig. 5.1 shows an example of a block matrix with blocks B_1 and B_2 fulfilling constraint 1 and the corresponding check of constraint 2 in the template structure. In this example, **d** has to be at least 2, if blocks B_1 and B_2 are to be connected by an edge in the DAG.

In the DAG we then weigh each node with the normalized score (as defined in Equ. (5.8)) of the corresponding block. Then Procedure 2 identifies the optimal path in the DAG. This will be explained in more detail in the following paragraph.

The idea of the algorithm is as follows. Beginning from the artificial start node, we perform a depth first search (*DFS*) in the DAG identifying the lowest scoring path according to $P(T', r)$. However, $P(T', r)$ is not only dependent on the summed score of the blocks but also on r, the length of the path (number of nodes in the path $\equiv$ number blocks in the alignment). When reaching a node v in the DFS, it is impossible to determine the overall best successor for this node. However, when the number of allowed successors of v in the path is fixed, the solution may be found. Therefore, we do not save a single best successor for each

Figure 5.1: A block matrix with a gap of length d in the target sequence highlighted and the corresponding gap in the structure of the template.





Procedure 2 DFS which fills the array succ

```
/* adj     ... adjacency list of the nodes in the graph
   visited ... array of boolean variables saving the DFS status of each node
   v       ... current node looked at in the graph
   succ    ... two dimensional array saving the optimal successors          */

def SimShiftDB_DFS(adj,visited,v,succ)
 1: best_succ ← [] /*empty array*/
 2: for w in adj[v] do
 3:    if visited[w] = 0 then
 4:       SimShiftDB_DFS(adj,visited,w,succ)
 5:    /*merge arrays best_succ and succ[w] favoring higher scoring paths*/
 6:    best_succ ← merge(best_succ, succ[w])
 7: visited[v] ← 1
 8: for k in best_succ do
 9:    succ[v][k+1] ← best_succ[k]
```

node, but we keep an array of optimal successors, for each possible number of succeeding blocks, named *succ* in Procedure 2 (succ[v][3], for example, saves the optimal successors of v given that v is first node in a path consisting of three blocks). After the DFS finishes, succ[start] contains a list of pointers, pointing to the start node of the optimal path for each possible path length. Finally, we select the combination of blocks (path) achieving the lowest p-value from succ[start]. The worst case running time of Procedure 2 is $O(e * (n + m))$ with e being the number of edges in the DAG and n and m being the length of the target and the length of the template, respectively. Note that the DAG is sparsely connected and therefore in practice e is in the order of $(n + m)^2$.

To give a simple estimation of SimShiftDB's running time in practice, we apply the algorithm to our set of target proteins (average length of 122 residues) and calculate the average time per protein. In our implementation one database[1] search on an standard laptop (Intel T2500, 2.0 GHz, 1 GB RAM) takes about 10 minutes. By discarding longer blocks in Step 1 of the algorithm the running time may be strongly decreased. Discarding all blocks with a length less than 10, for example, results in a strong running time decrease to about 1 minute per protein.

[1] 64839 proteins with an average length of 183 residues

5.6 Results

5.6.1 Evaluation of the Modeling Performance

To evaluate the performance of our algorithm, we compare our method to 123D, an established threading method, and the standard tool used by spectroscopists working with chemical shifts, namely TALOS. Our target set has to fulfill two constraints:

- The chemical shift data shall be of high quality (not corrupted by errors as noted by Zhang et al. [2003]).
- Chemical shifts for ^{1}H, $^{13}C_{\alpha}$, $^{13}C_{\beta}$ and $^{13}C'$ have to be available, as the substitution score is calculated based on these values.

As it is often hard to check the reliability of chemical shift data, we use a set of six target structures which were provided by the group of Prof. Dr. Horst Kessler from the Technische Universität München. The data for PH1500-N (unpublished), HAMP [Hulko et al., 2006] , PHS018 [Coles et al., 2006], KDP [Haupt et al., 2004] and VAT-N [Coles et al., 1999] was measured directly by this group. The data for JOSEPHIN [Nicastro et al., 2005], which was solved by a different group, was checked for its correctness. As all of these structures were recently resolved, three dimensional data is also available. This way we can reliably check the quality of our predictions. The set-up of our experiment is as follows:

- It is required that all methods give torsion angle prediction for at least 80% of the target protein. For 123D and SimShiftDB, we sort the alignments produced by the quality score of the respective method (alignment score for 123D and e-value for SimShiftDB) and take as many alignments (starting from the best) as necessary such that at least 80% of the residues of the target protein have an assigned residue from a template structure. The assignment is done favoring alignments with better score if a residue from the target structure is assigned multiple times in separate alignments. Concerning the comparison of SimShiftDB to 123D, we additionally discard all alignments with sequence identity $\geq 90\%$ to remove trivial solutions. As TALOS predicts backbone torsion angles for all residues of the target, no additional work is required in this case.
- To evaluate the torsion angle predictions, we build a model for the torsion angles from the target structure (using the torsion angles of the assigned residues from the templates). Then we calculate the torsion angles for our target using STRIDE [Frishman and Argos, 1995]. Now it is possible to assess the average error in torsion angle prediction by using the STRIDE calculations as a Standard of Truth.

Fig. 5.2 and Fig. 5.3 show the percentage of torsion angles per structure where SimShiftDB outperforms 123D and TALOS (for Φ and Ψ angles, respectively). SimShiftDB outperforms 123D in 62% and 69% of all cases and 35% and 36% of all backbone torsion angles predicted by SimShiftDB have a smaller error than those predicted by TALOS. To check that the difference between TALOS and SimShiftDB is not just marginal, we calculate the mean error of both methods for the cases where SimShiftDB outperforms TALOS (see Fig. 5.4 and 5.5 for Details). SimShiftDB reduces the error (compared to TALOS) by more than 60%.

5.6.2 Evaluation of the P-Value Correctness

Two sets, S_1 and S_2, of random chemical shift alignments are constructed as follows. Step 1 of the SimShiftDB algorithm is performed for each target-template-pair. Based on the identified blocks, two DAGs, namely G_1 and G_2, are build. In G_2 nodes which fulfill constraints 1 and 2 (see page 52) are connected, whereas in G_1 constraint 1 has to be fulfilled only. Then ten nodes are drawn from each DAG without replacement. For every node n, we construct a random path in the DAG starting in n until we reach a node with outdegree zero. Each prefix of the path in G_1 (or G_2) yields an alignment, which is added to S_1 (or S_2, respectively).

Using the procedure decribed above, we are able to construct random alignments which are built of up to seven blocks. For each constructed alignment from S_1 or S_2, we calculate the empirical p-value and compare it to the theoretical p-value. The results of this comparison for alignments consisting of one to seven blocks are shown in Appendix B and C, for S_1 and S_2, respectively. The empirical p-value is always less than the theoretical p-value for both sets. Therefore, the theoretical p-value provides a conservative estimate, both in theory and practice.

5.7 Discussion

We developed a method which builds models for target proteins of unknown structure using chemical shift data measured in NMR Spectroscopy. The method has been evaluated on a small, but very reliable test set. From the results presented in the last section, we draw the following conclusions:

- SimShiftDB strongly outperforms methods which are based on amino acid sequence information alone and should therefore be used whenever chemical shift data is available.
- When comparing to TALOS, both methods show their strength. However, SimShiftDB is able to outperform TALOS in a significant number of cases.

Figure 5.2: Percentage of Φ-angle prediction where SimShiftDB outperforms 123D and TALOS

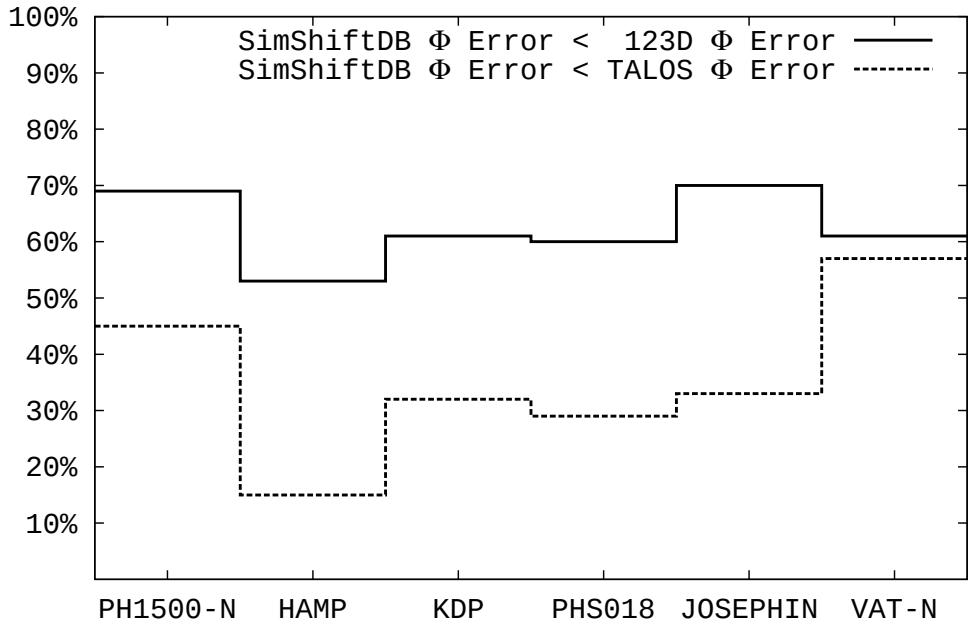


Figure 5.3: Percentage of Ψ-angle predictions where SimShiftDB outperforms 123D and TALOS

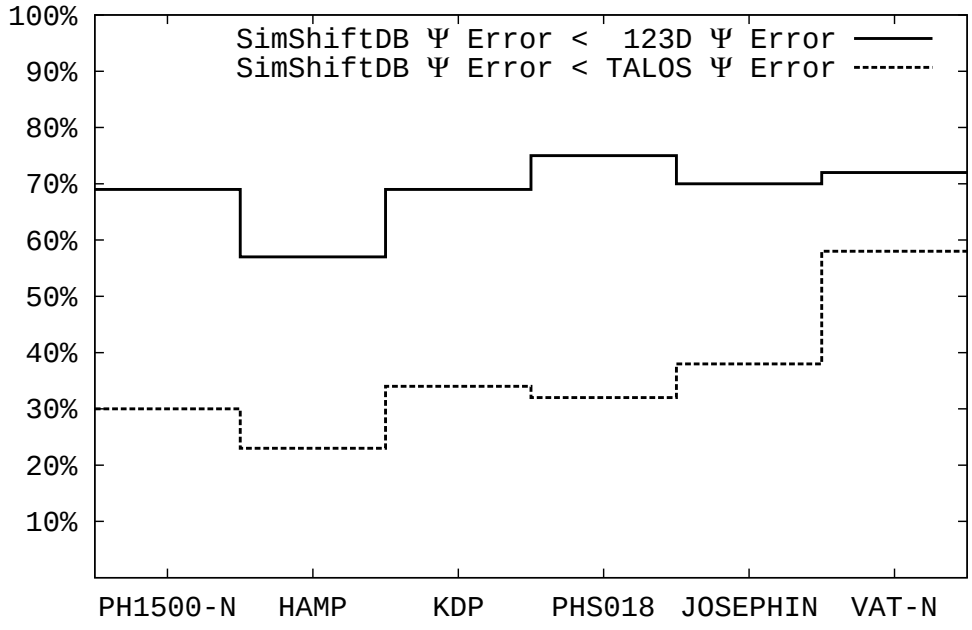


Figure 5.4: Φ-angle error compared to STRIDE for those prediction where SimShiftDB outperforms TALOS

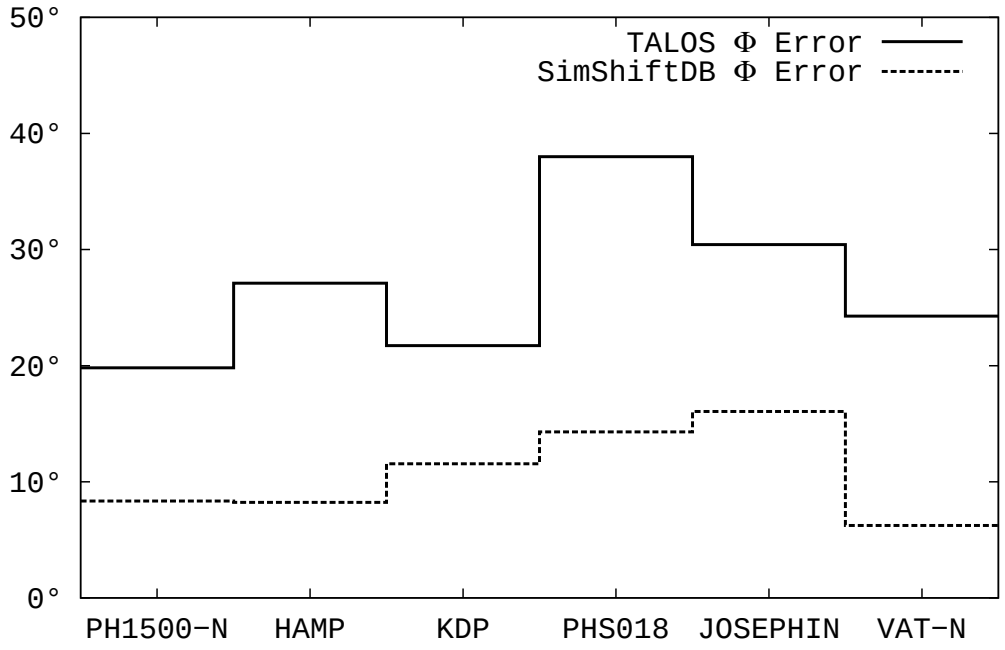
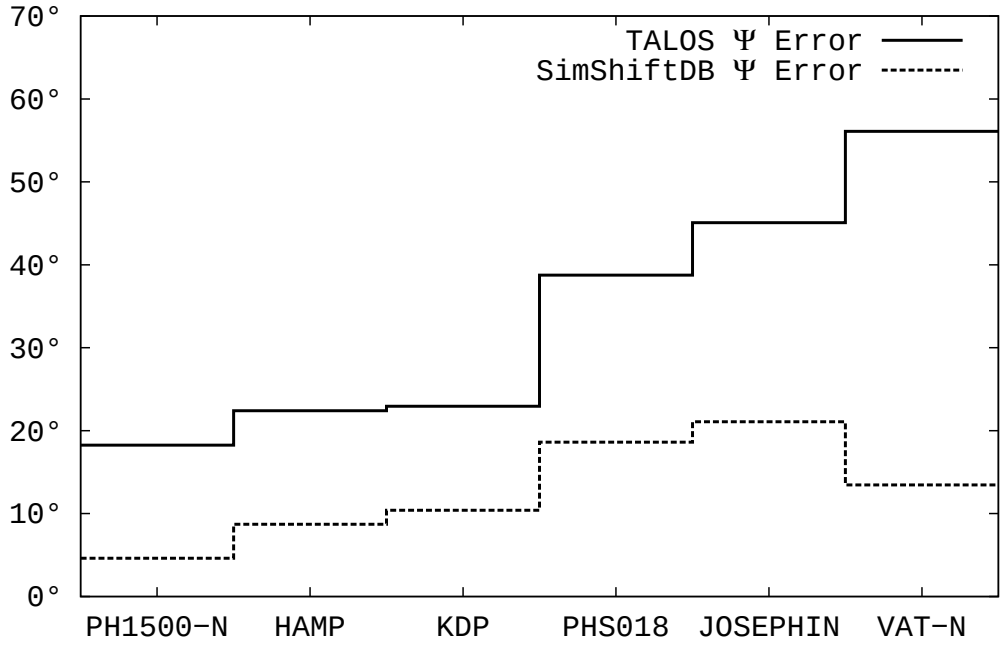


Figure 5.5: Ψ-angle error compared to STRIDE for those prediction where SimShiftDB outperforms TALOS



As also described in Chapter 3 chemical shift data is often inconsistently referenced, which may strongly influence the quality of any further data analysis. Many approaches as TALOS or PSSI[2] [Wang and Jardetzky, 2002] use simple re-referencing protocols which are applied before the actual method. In the next chapter, we present the *Chemical Shift Pipeline*, a combination of CheckShift and SimShiftDB. Using this combination, wrongly referenced chemical shifts data does not hamper the quality of SimShiftDB's results any more. Therefore, we are now able to define a larger benchmark set based on entries from the BMRB database. We also calculate a set of different chemical shift substitution matrices and propose a "best" matrix based on the results for the proteins in the benchmark set.

5.8 Availability

`http://shifts.bio.ifi.lmu.de/`

[2] A method which calculates secondary structure based on chemical shifts

6 The Chemical Shift Pipeline

6.1 Introduction

In this chapter, the main result of this work, the so-called *Chemical Shift Pipeline*, is presented. The Chemical Shift Pipeline combines CheckShift and SimShiftDB, to increase the performance of the homology modeling procedure, making it insensitive to inconsistent referencing. Additionally, the SimShiftDB algorithm is extended, being able to cope with missing chemical shift values in a better way.

As seen in Chapter 5, some parameters have to be defined for SimShiftDB to run, the most influential being the chemical shift substitution matrix. Here we compile a set of 17 chemical shift substitution matrices to test whether the performance of the SimShiftDB may be improved, by using the "right" substitution matrix. It is always an issue to define a measure of correctness for chemical shift alignments. Therefore, a benchmark set of 144 (sufficiently large) proteins is derived, where both chemical shifts and structural data are available. Subsequently, different parameter combinations are evaluated on the benchmark set. Finally the performance of the different parameter settings is presented and the "best" setting is proposed.

6.2 Coping with Missing Chemical Shift Data

As for all data from the laboratory, chemical shift data may not only be error-prone (see Chapter 3), but also simply missing. It is always a difficult question, how to cope with missing data. Especially when working with chemical shifts, it seems worthwhile spending some time on this issue. To give an example, if the C_α and C_β shift are given and only the C' shift is missing, it is somewhat uneconomic to mark the whole residue as unusable. Therefore, a different alphabet for amino acids with associated chemical shifts is defined, where chemical shifts may not only be *week* or *strong*, but also *missing*. A residue with a missing C' shift is now not unusable, but induces only a different lookup in the chemical shift substitution matrix. As the chemical shifts substitution matrices are built from different sets of reference alignments, one has to keep in mind that the larger the alphabet gets, the worse is the statistical conditioning. Therefore, we do not allow letters in out alphabet in which more than three chemical shift values are available.

Residue Pair:

C	**C_β**	**C_α**	**N**	**H**	**H_α**			**H_α**	**H**	**N**	**C_α**	**C_β**	**C**
2	0	1	1	1	0	**L**	**A**	0	0	1	2	0	1

Highest Priority Mask (Number 39):

H_α	H	N	C_α	C_β	C
✘	✘			✘	

Masked Residues:

C_β	**H**	**H_α**			**H_α**	**H**	**C_β**
0	1	0	**L**	**A**	0	0	0

Matrix Lookup:

$$\mathbf{1}_{90\%}[39][\texttt{010L-A000}] = -4$$

Figure 6.1: Example of the similarity score retrieval.

In practice masks are defined for every legal subset of the available chemical shift values (see Table 6.1). If two residues $\mathcal{A}$ and $\mathcal{B}$ shall be compared, the highest priority mask, which may be applied to both $\mathcal{A}$ and $\mathcal{B}$ (thereby not masking shift values that are missing), is identified. Now the chemical shift substitution matrix for this mask is selected and the similarity score is retrieved (see Figure 6.1 for an example). The chemical shift priority is defined as follows: C_α has the highest priority, followed by C_β, C', H_α, H and N. These priorities were selected based on the experience of several researchers working in NMR Spectroscopy. Based on the constraints imposed, it is not necessary to calculate and fill a matrix containing $(20 \cdot 3^6)^2 = 212.576.400$ elements, however, we need only $\binom{6}{3} + \binom{6}{2} + \binom{6}{1} = 20 + 15 + 6 = 41$ matrices (one for every legal mask) of size $(20 \cdot 2^3)^2 = 25.600$, which multiplies to a total of 1.049.600 matrix elements.

6.3 Chemical Shift Substitution Matrices

As described in Section 5.3, the chemical shift substitution matrices are calculated based on a set of protein-proteins alignments which are assumed to be correct. Various source for alignments exist. Three sets of alignments were selected based

Table 6.1: Masks used for selection of certain chemical shift values.

	H_α	H	N	C_α	C_β	C
1						✘
2					✘	
3					✘	✘
4				✘		
5				✘		✘
6				✘	✘	
7				✘	✘	✘
8			✘			
9			✘			✘
10			✘		✘	
11			✘		✘	✘
12			✘	✘		
13			✘	✘		✘
14			✘	✘	✘	
15		✘				
16		✘				✘
17		✘			✘	
18		✘			✘	✘
19		✘		✘		
20		✘		✘		✘
21		✘		✘	✘	
22		✘	✘			
23		✘	✘			✘
24		✘	✘		✘	
25		✘	✘	✘		
26	✘					
27	✘					✘
28	✘				✘	
29	✘				✘	✘
30	✘			✘		
31	✘			✘		✘
32	✘			✘	✘	
33	✘		✘			
34	✘		✘			✘
35	✘		✘		✘	
36	✘		✘	✘		
37	✘	✘				
38	✘	✘				✘
39	✘	✘			✘	
40	✘	✘		✘		
41	✘	✘	✘			

on an extensive literature research. All three are tested as possible basis for the chemical shifts substitution matrices:

1. A set of protein family alignments where families correspond to the protein domain sets as defined by the SCOP [Murzin et al., 1995] classification.
2. A set of multiple alignments of protein families, defined through structural similarity calculated using the CE [Shindyalov and Bourne, 1998] algorithm.
3. The S4 set of alignments [Casbon and Saqi, 2005] which consists of multiple alignments of proteins classified as being in the same superfamily as according to the SCOP classification.

Sets **1** and **2** are available from the DMAPS database [Guda et al., 2006].

By further restricting sets **1** and **2** to include only alignments which do not excess as certain sequence identity, 17 sets were compiled from these three data sources (for set **3** this restriction is not needed, as according to [Casbon and Saqi, 2005] all alignments have $\leq 40\%$ equal residues). Thereby the following values were used as cutoffs for the maximal sequence identity: 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%. Now for every set we calculate the chemical shift substitution matrices in the following way:

1. Minimal and maximal shifts for each atom and residue type are retrieved from the template database described in Section 5.2.
2. For every amino acid and every atom type, we divide the range between the minimal and the maximal shift observed into two parts. Every shift associated to a residue in a reference alignment is then converted into a integer value, being either *weak* (`0`) (situated in the lower part of the range), *strong* (`1`) (situated in the upper part of the range), or *missing* (`2`). This discrete representation of chemical shifts together with the amino acid type is then used as a letter in our sequence alphabet (see Table 6.2 for an example).
3. Then the relative frequencies of each pair of letters $\mathcal{A}, \mathcal{B}$ ($q_{\mathcal{A},\mathcal{B}}$) are calculated from the reference alignments (see Figure 6.2). For each family of the respective alignment sets, we add pseudo counts, thereby removing the bias introduced by the over-representation of certain protein families.
4. The relative frequencies of each letter $\mathcal{A}$ occurring independently ($p_{\mathcal{A}}$) are derived from the back calculated chemical shifts in the template database.
5. Finally log-odds scores are calculated for every combination of two letters $\mathcal{A}$ and $\mathcal{B}$:

$$o_{\mathcal{A},\mathcal{B}} = \log\left(\frac{q_{\mathcal{A},\mathcal{B}}}{e_{\mathcal{A},\mathcal{B}}}\right), \tag{6.1}$$

where

$$e_{\mathcal{A},\mathcal{B}} = \begin{cases} 2 * p_{\mathcal{A}} * p_{\mathcal{B}} & \text{if} \quad \mathcal{A} \neq \mathcal{B}, \\ p_{\mathcal{A}}^2 & \text{otherwise.} \end{cases} \tag{6.2}$$

The log-odds scores are multiplied with a normalization factor η and rounded to the nearest integer. The shift substitution matrix entries $s_{\mathcal{A},\mathcal{B}}$ are then formally defined as

$$s_{\mathcal{A},\mathcal{B}} = \lfloor o_{\mathcal{A},\mathcal{B}} * \eta + 0.5 \rfloor. \tag{6.3}$$

Here we set the value of η equal to 10. This value was chosen based on a thorough inspection of the values $o_{\mathcal{A},\mathcal{B}}$, trying to sacrifice as little information as possible.

Note that steps 3, 4 and 5 are executed for every mask shown in Table 6.1.

6.4 The Benchmark Set

To test the performance of the Chemical Shift Pipeline, a benchmark set has to be defined for which both chemical shifts and the three-dimensional structure of the protein are available. The BMRB [Seavey et al., 1991] is the main public repository for chemical shift data[1]. However, there is no consistent mapping to PDB, therefore making it difficult to relate structural with chemical shift information reliably. Therefore, a mapping between BMRB and ASTRAL is calculated based on amino acid sequence similarity. Every entry in the benchmark set has to fulfill the following constraints.

- A 100% sequence match to an ASTRAL entry.
- At least 100 residues with associated chemical shifts, to exclude very short protein fragments (e.g., single helices)

The mapping procedure used here is the same as described in Section 4.2.1. Based on these constraints a set of 144 chemical shift set was derived. The benchmark set is listed in Appendix D. For each member we show the corresponding ASTRAL and BMRB identifier, its secondary structure composition and its length.

6.5 Results

To evaluate the performance of the different chemical shift substitution matrices, the chemical shift pipeline is applied to all entries in the benchmark set: At first all

[1]The snapshot of the BMRB used here was taken on the 11$^{\text{th}}$ of June, 2007

Table 6.2: Reference alignment with associated shifts (floating point value and their discrete equivalents).

1ar5b **EKDLAFNLAGHVNHSVFWKNMAP**
1qnna **EGGIFNNAGQTLNHNLYFTQFRP**

C	C_β	C_α	N	H	H_α			H_α	H	N	C_α	C_β	C
Floating point values:													
175.2	33.6	55.0	119.8	8.0	4.5	**E**	**E**	4.1	8.5	118.7	59.0	29.3	179.2
176.1	n/a	46.2	109.0	8.7	4.1	**G**	**K**	4.1	7.9	120.0	59.4	32.3	178.6
175.2	n/a	48.0	105.2	8.8	3.7	**G**	**D**	4.5	8.5	119.6	57.1	40.6	178.6
178.1	37.2	64.2	121.5	7.7	3.7	**I**	**L**	3.9	8.2	120.6	58.1	42.2	178.8
176.8	39.2	60.7	122.5	7.8	4.0	**F**	**A**	4.0	8.0	120.7	55.1	18.3	179.3
177.2	38.4	56.6	116.3	8.4	4.1	**N**	**F**	4.1	8.6	119.4	61.7	39.3	n/a
177.5	38.6	56.0	117.2	n/a	4.2	**N**	**N**	4.2	8.2	116.0	56.1	38.4	177.3
179.9	18.4	54.9	122.5	8.0	3.9	**A**	**L**	3.7	8.2	121.9	58.0	42.1	178.9
176.0	n/a	47.9	104.6	7.9	3.0	**G**	**A**	4.0	8.0	120.9	n/a	18.2	179.9
179.0	28.5	58.5	120.6	7.5	3.8	**Q**	**G**	2.6	7.4	104.4	47.7	n/a	175.5
176.4	68.6	66.5	116.9	7.6	3.9	**T**	**H**	4.3	8.1	120.8	59.0	30.1	177.5
178.4	42.0	58.2	122.3	7.9	3.9	**L**	**V**	3.4	7.9	120.4	66.1	31.4	177.7
177.2	38.5	57.1	116.7	8.3	4.2	**N**	**N**	4.1	8.1	116.4	56.8	38.4	177.2
177.7	28.5	58.7	117.1	8.3	4.0	**H**	**H**	3.6	7.9	117.4	58.5	28.6	178.1
177.4	38.2	56.5	118.5	8.4	4.5	**N**	**S**	4.3	8.0	117.2	61.6	62.6	176.9
178.7	41.9	57.5	119.2	7.8	4.1	**L**	**V**	3.6	7.6	120.6	65.7	31.1	177.6
178.2	38.6	60.9	119.7	8.1	4.0	**Y**	**F**	4.2	7.7	119.5	61.3	39.2	176.9
176.4	39.2	60.5	117.8	8.4	4.6	**F**	**W**	4.8	7.7	118.2	60.1	28.7	178.8
175.1	69.0	62.1	107.3	7.6	4.4	**T**	**K**	4.2	7.6	116.2	59.1	31.6	178.6
175.4	29.3	56.3	119.7	7.3	4.2	**Q**	**N**	4.6	8.1	115.2	54.0	38.7	174.5
175.2	41.3	56.4	117.1	7.5	5.2	**F**	**M**	4.8	6.9	119.1	53.6	35.3	174.6
173.5	33.9	54.3	120.6	8.1	4.6	**R**	**A**	4.3	7.8	120.6	50.5	21.9	174.7
176.6	32.1	63.0	n/a	n/a	4.3	**P**	**P**	4.2	n/a	n/a	62.2	31.9	176.7
Associated discrete values:													
0	1	0	0	1	1	**E**	**E**	0	1	0	1	0	1
1	0	1	0	1	1	**G**	**K**	0	0	0	1	0	1
1	0	1	0	1	0	**G**	**D**	1	1	0	1	0	1
1	0	1	1	0	0	**I**	**L**	0	1	1	1	0	1
1	0	1	1	0	0	**F**	**A**	0	0	0	1	0	1
1	0	1	0	1	0	**N**	**F**	0	1	0	1	0	2
1	0	1	0	2	1	**N**	**N**	1	1	0	1	0	1
1	0	1	1	0	0	**A**	**L**	0	1	1	1	0	1
1	0	1	0	0	0	**G**	**A**	0	0	0	2	0	1
1	0	1	1	0	0	**Q**	**G**	0	0	0	1	0	1
1	0	1	1	0	0	**T**	**H**	1	0	1	1	0	1
1	0	1	1	0	0	**L**	**V**	0	0	1	1	0	1
1	0	1	0	1	1	**N**	**N**	0	1	0	1	0	1
1	0	1	0	0	0	**H**	**H**	0	0	0	1	0	1
1	0	1	0	1	1	**N**	**S**	1	0	0	1	0	1
1	0	1	0	0	1	**L**	**V**	0	0	1	1	0	1
1	0	1	0	1	0	**Y**	**F**	1	0	0	1	0	1
1	0	1	0	1	1	**F**	**W**	1	0	0	1	0	1
1	0	0	0	0	1	**T**	**K**	1	0	0	1	0	1
0	0	0	0	0	1	**Q**	**N**	1	1	0	1	0	0
1	1	1	0	0	1	**F**	**M**	1	0	0	0	1	0
0	1	0	1	0	1	**R**	**A**	1	0	0	0	1	0
0	0	0	0	0	1	**P**	**P**	1	0	0	0	0	0

Mask (Number 21):

H_α	H	N	C_α	C_β	C
	✘		✘	✘	

Aligned Residues:

C	**C_β**	**C_α**	**N**	**H**	**H_α**			**H_α**	**H**	**N**	**C_α**	**C_β**	**C**
1	0	1	1	0	0	**F**	**A**	0	0	0	1	0	1
1	0	1	1	1	0	**L**	**A**	0	0	1	1	0	1
1	0	1	0	1	0	**N**	**F**	0	1	0	1	0	**2**
1	0	1	0	**2**	1	**N**	**N**	1	1	0	1	0	1
1	0	1	1	0	0	**A**	**L**	0	1	1	1	0	1
1	0	1	0	0	0	**G**	**A**	0	0	0	**2**	0	1
1	0	1	1	0	0	**Q**	**G**	0	0	0	1	0	1

Masked Aligned Residues:

C_β	**C_α**	**H**			**H**	**C_α**	**C_β**
0	1	0	**F**	**A**	0	1	0
0	1	1	**L**	**A**	0	1	0
0	1	1	**N**	**F**	1	1	0
0	1	0	**A**	**L**	1	1	0
0	1	0	**Q**	**G**	0	1	0

Relative frequencies:

$$q_{21}(\textbf{010A-F010}) = \frac{1}{5}$$
$$q_{21}(\textbf{010F-N110}) = \frac{1}{5}$$
$$q_{21}(\textbf{010A-L110}) = \frac{2}{5}$$
$$q_{21}(\textbf{010G-Q010}) = \frac{1}{5}$$

Figure 6.2: Example of the calculation of pairwise relative frequencies for mask number 21.

targets are re-referenced using CheckShift (see Chapter 3). Then SimShiftDB is run for each target. This is repeated with all chemical shift substitution matrices calculated. All results were calculated 6 times with the minimal block length parameter varying from 5 to 10. However, as there were essentially no differences, for the sake of simplicity, we only present the results for the minimal block length set to 10. Subsequently, we analyze all alignments which achieve an e-value of at most 10^{-3}. Then we use these alignments to infer torsion angles for the target from the associated residues of the template. It is extremely interesting to evaluate the performance of SimShiftDB also in cases where sequence similarity is low. Therefore, 8 evaluations were performed, for each of which a different maximum sequence similarity in the evaluated alignment was defined. Alignments exceeding the maximum sequence similarity were excluded from the evaluation.

Some notations are defined for the presentation of the results:

- Torsion angles with an error of less than or equal to 15° are considered correct (marked as $\checkmark_\Phi$, $\checkmark_\Psi$ in the tables).
- Torsion angles with an error of more than 30° are considered completely wrong (marked as $\boldsymbol{\times}_\Phi$, $\boldsymbol{\times}_\Psi$ in the tables).
- Sequence identity is defined as the percentage of identical residues **in** the alignment. Therefore, the number of aligned residue pairs which are identical is derived and divided by the number of all amino acid pairs in the alignment.
- The rows in the result tables correspond to different substitution matrices. The number relates to the associated alignment set (as listed in Section 6.3). The subscript describes the maximal sequence identity in the alignments used to calculate the pairwise frequencies. For example the row marked with $\mathbf{1}_{90\%}$ gives the results for the matrix calculate from the SCOP alignments taken from the DMAPS database with a maximal sequence identity of 90%.

In Tables E.1 to E.8, the results of the evaluation of the Chemical Shift Pipeline on the benchmark set are given. The rows in the tables are ordered by the percentage of correct Φ angle predictions (which correlates strongly with Ψ angle correctness). The column **C** give the percentage of the number of target residues for which a prediction has been made. In the columns Δ_Φ, Δ_Ψ the averaged error of **all** torsion angle predictions is shown. From this evaluation, we propose to use the matrix $\mathbf{1}_{90\%}$ (highlighted with a gray background) as it gives a good tradeoff between sensitivity and specificity.

What is also interesting is the performance of the Chemical Shift Pipeline on different types of secondary structure. As the three dimensional structure of all proteins in the benchmark set is known, every target residue can be classified as being part of on of the three secondary structure states: **H** - Helix, **S** -

Sheet, **C** - Coil. The secondary structure for the target is calculated again using STRIDE [Frishman and Argos, 1995], and the 5 letter secondary structure code produced is subsequently converted to the three states, by defining **G**, **I** and **H** as helix, **B** and **E** as sheet, and everything else as coil. In Table 6.3 the secondary structure percentage in correct and completely wrong predictions is listed for every maximal sequence identity in the alignments. As the secondary structure prior in the test set is 38%, 24%, and 37%, for helix, sheet, and coil, respectively, one can say that the predictions for sheet match well with the percentage induced by the bias of the benchmark set. What can also be derived is that the percentage of correct predicions in helix increases, whereas the corresponding percentage in coil regions decreases. This seems very logical, as coil regions as often measured incorrectly and are definitly harder to predict than the very well structured helix regions. This test proves empirically, that the Chemical Shift Pipeline has no bias when comparing helix to sheet performance.

6.6 Discussion

What has been presented is a new way to analyze chemical shift data, leading to the creation of a three dimensional model for a target protein at an early stage of the NMR experiments. It has been shown that the Chemical Shift Pipeline produces high quality alignments, even in cases where sequence similarity is low. Using the e-value as a tool to separate the wheat from the chaff results in highly reliable predictions. We are convinced that the availability of the Chemical Shift Pipeline will support researchers in NMR spectroscopy, thereby significantly speeding up the structure solving process.

6.7 Availability

`http://shifts.bio.ifi.lmu.de/`

Table 6.3: Secondary structure percentage in correct and completely wrong predictions for $\mathbf{1}_{90\%}$. The prior percentage of secondary structure in the benchmark set is 38%, 24%, 37%, for helix, sheet and coil, respectively. The first column ($\leq_{\text{Id.}}$) gives the maximal sequence identity for the evaluated alignments. In the upper part we show the secondary structure percentage of correct predictions ($\checkmark_\Phi$, $\checkmark_\Psi$) in the lower part of the table the corresponding percentages for completely wrong predictions ($\boldsymbol{\times}_\Phi$, $\boldsymbol{\times}_\Psi$) are shown.

$\leq_{\text{Id.}}$	$\checkmark_\Phi$H	$\checkmark_\Phi$S	$\checkmark_\Phi$C	$\checkmark_\Psi$H	$\checkmark_\Psi$S	$\checkmark_\Psi$C
30%	64%	21%	15%	63%	23%	14%
40%	56%	26%	18%	55%	27%	18%
50%	54%	25%	21%	53%	27%	21%
60%	51%	25%	24%	50%	27%	24%
70%	49%	25%	26%	48%	26%	26%
80%	48%	25%	27%	47%	27%	26%
90%	48%	24%	27%	47%	26%	26%
100%	46%	25%	29%	46%	26%	29%
	$\boldsymbol{\times}_\Phi$H	$\boldsymbol{\times}_\Phi$S	$\boldsymbol{\times}_\Phi$C	$\boldsymbol{\times}_\Psi$H	$\boldsymbol{\times}_\Psi$S	$\boldsymbol{\times}_\Psi$C
30%	17%	31%	52%	21%	27%	52%
40%	15%	32%	53%	17%	29%	54%
50%	13%	31%	56%	16%	28%	56%
60%	14%	29%	57%	16%	26%	58%
70%	14%	29%	58%	16%	25%	59%
80%	15%	28%	57%	17%	25%	59%
90%	15%	28%	58%	16%	25%	59%
100%	14%	28%	57%	16%	25%	59%

7 Outlook

What has been presented is a new way to identify structural homologues for proteins of interest, which is solely based on analyzing amino acid sequence and chemical shift data. It has been shown that SimShiftDB, given an additional automatic reference correction using CheckShift, achieves highly accurate predictions. What remains is the question: Where to go from now on?

If we were to define the ultimate goal of analyzing protein chemical shifts, the answer gives us a 'déjà écouté'-feeling. As many people who analyze the amino acid sequence of proteins, we want to find a way to get straight from the chemical shifts to the three-dimensional structure. Is this realistic? Definitely not in a general sense! However, we are convinced that using SimShiftDB the range of structures which can be modeled is strongly expanded. To get a high quality structure out of a SimShiftDB database search, structural constraints, extracted from several alignments, have to be combined. Subsequently, the resulting crude model has to be refined to yield a realistic three dimensional protein structure. Finding the best combination and refinement procedures is exactly where further research is headed.

Bibliography

N. N. Alexandrov, R. Nussinov, and R. M. Zimmer. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac Symp Biocomput*, pages 53–72, 1996.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990. doi: 10.1006/jmbi.1990.9999.
URL `http://dx.doi.org/10.1006/jmbi.1990.9999`.

S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

D. Braun, G. Wider, and K. Wuethrich. Sequence-Corrected 15N" Random Coil" Chemical Shifts. *Journal of the American Chemical Society*, 116(19):8466–8469, 1994.

A. Buckingham. Chemical shifts in the nuclear magnetic resonance spectra of molecules containing polar groups. *Canadian Journal of Chemistry*, 38(2): 300–307, 1960.

J. Casbon and M. A. S. Saqi. S4: structure-based sequence alignments of scop superfamilies. *Nucleic Acids Res*, 33(Database issue):D219–D222, Jan 2005. doi: 10.1093/nar/gki043.
URL `http://dx.doi.org/10.1093/nar/gki043`.

J.-M. Chandonia, G. Hon, N. S. Walker, L. L. Conte, P. Koehl, M. Levitt, and S. E. Brenner. The ASTRAL compendium in 2004. *Nucleic Acids Res*, 32 (Database issue):D189–D192, Jan 2004. doi: 10.1093/nar/gkh034.
URL `http://dx.doi.org/10.1093/nar/gkh034`.

M. Coles, T. Diercks, J. Liermann, A. Groger, B. Rockel, W. Baumeister, K. K. Koretke, A. Lupas, J. Peters, and H. Kessler. The solution structure of VAT-N

reveals a 'missing link' in the evolution of complex enzymes from a simple $\beta\alpha\beta\beta$ element. *Curr Biol*, 9(20):1158–1168, Oct 1999. doi: 10.1016/S0960-9822(00)80017-2.
URL http://dx.doi.org/10.1016/S0960-9822(00)80017-2.

M. Coles, M. Hulko, S. Djuranovic, V. Truffault, K. Koretke, J. Martin, and A. N. Lupas. Common evolutionary origin of swapped-hairpin and double-psi beta barrels. *Structure*, 14(10):1489–1498, Oct 2006. doi: 10.1016/j.str.2006.08.005.
URL http://dx.doi.org/10.1016/j.str.2006.08.005.

G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR*, 13(3):289–302, Mar 1999.

S. R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

D. Fischer, A. Elofsson, L. Rychlewski, F. Pazos, A. Valencia, B. Rost, A. R. Ortiz, and R. L. Dunbrack. Cafasp2: the second critical assessment of fully automated structure prediction methods. *Proteins*, Suppl 5:171–183, 2001.

P. Fontana, E. Bindewald, S. Toppo, R. Velasco, G. Valle, and S. C. E. Tosatto. The SSEA server for protein secondary structure alignment. *Bioinformatics*, 21(3):393–395, Feb 2005. doi: 10.1093/bioinformatics/bti013.
URL http://dx.doi.org/10.1093/bioinformatics/bti013.

D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579, Dec 1995. doi: 10.1002/prot.340230412.
URL http://dx.doi.org/10.1002/prot.340230412.

S. W. Ginzinger and J. Fischer. SimShift: Identifying structural similarities from NMR chemical shifts. *Bioinformatics*, 22(4):460–465, Feb 2006. doi: 10.1093/bioinformatics/bti805.
URL http://dx.doi.org/10.1093/bioinformatics/bti805.

S. W. Ginzinger, T. Gräupl, and V. Heun. SimShiftDB: Chemical-shift-based homology modeling. *Proceedings of the First Conference on Bioinformatics Research and Development, Lecture Notes in Bioinformatics*, 4414:357–370, 2007.

C. Guda, S. Lu, E. D. Scheeff, P. E. Bourne, and I. N. Shindyalov. Ce-mc: a multiple protein structure alignment server. *Nucleic Acids Res*, 32(Web Server issue):W100–W103, Jul 2004. doi: 10.1093/nar/gkh464.
URL http://dx.doi.org/10.1093/nar/gkh464.

C. Guda, L. Pal, I. Shindyalov, and O. Journals. DMAPS: a database of multiple alignments for protein structures. *Nucleic Acids Research*, 2006.

C. Haigh and R. Mallion. New tables of ring current shielding in proton magnetic resonance. *Org. Magn. Reson*, pages 203–228, 1972.

M. Haupt, M. Bramkamp, M. Coles, K. Altendorf, and H. Kessler. Inter-domain motions of the N-domain of the KdpFABC complex, a P-type ATPase, are not driven by ATP-induced conformational changes. *J Mol Biol*, 342(5):1547–1558, Oct 2004. doi: 10.1016/j.jmb.2004.07.060.
URL `http://dx.doi.org/10.1016/j.jmb.2004.07.060`.

M. Haupt, M. Bramkamp, M. Heller, M. Coles, G. Deckers-Hebestreit, B. Herkenhoff-Hesselmann, K. Altendorf, and H. Kessler. The holo-form of the nucleotide binding domain of the kdpfabc complex from escherichia coli reveals a new binding mode. *J Biol Chem*, 281(14):9641–9649, Apr 2006. doi: 10.1074/jbc.M508290200.
URL `http://dx.doi.org/10.1074/jbc.M508290200`.

M. Heinig and D. Frishman. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res*, 32 (Web Server issue):W500–W502, Jul 2004. doi: 10.1093/nar/gkh429.
URL `http://dx.doi.org/10.1093/nar/gkh429`.

M. Heller, M. John, M. Coles, G. Bosch, W. Baumeister, and H. Kessler. Nmr studies on the substrate-binding domains of the thermosome: structural plasticity in the protrusion region. *J Mol Biol*, 336(3):717–729, Feb 2004. doi: 10.1016/j.jmb.2003.12.035.
URL `http://dx.doi.org/10.1016/j.jmb.2003.12.035`.

S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci*, 89(22):10915–10919, Nov 1992.

L. Holm and C. Sander. Mapping the protein universe. *Science*, 273(5275): 595–603, Aug 1996.

M. Hulko, F. Berndt, M. Gruber, J. U. Linder, V. Truffault, A. Schultz, J. Martin, J. E. Schultz, A. N. Lupas, and M. Coles. The hamp domain structure implies helix rotation in transmembrane signaling. *Cell*, 126(5):929–940, Sep 2006. doi: 10.1016/j.cell.2006.06.058.
URL `http://dx.doi.org/10.1016/j.cell.2006.06.058`.

D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2):195–202, Sep 1999. doi: 10.1006/jmbi.1999.3091.
URL `http://dx.doi.org/10.1006/jmbi.1999.3091`.

D. Joseph, J. Meidanis, and P. Tiwari. Determining DNA Sequence Similarity Using Maximum Independent Set Algorithms for Interval Graphs. *Proceedings*

of the Third Scandinavian Workshop on Algorithm Theory, Lecture Notes in Bioinformatics, pages 326–337, 1992.

W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, 34(5):827–828, 1978.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12): 2577–2637, Dec 1983. doi: 10.1002/bip.360221211.
URL `http://dx.doi.org/10.1002/bip.360221211`.

S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci*, 87(6):2264–2268, Mar 1990.

S. Karlin and S. F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci*, 90(12):5873–5877, Jun 1993.

L. N. Kinch, J. O. Wrabl, S. S. Krishna, I. Majumdar, R. I. Sadreyev, Y. Qi, J. Pei, H. Cheng, and N. V. Grishin. Casp5 assessment of fold recognition target predictions. *Proteins*, 53 Suppl 6:395–409, 2003. doi: 10.1002/prot.10557.
URL `http://dx.doi.org/10.1002/prot.10557`.

M. H. Levitt. *Spin dynamcis: basics of nuclear magnetic resonance.* Wiley, Chichester, UK, 2001.

Y. Mao, F. Senic-Matuglia, P. P. D. Fiore, S. Polo, M. E. Hodsdon, and P. D. Camilli. Deubiquitinating function of ataxin-3: insights from the solution structure of the josephin domain. *Proc Natl Acad Sci*, 102(36):12700–12705, Sep 2005. doi: 10.1073/pnas.0506344102.
URL `http://dx.doi.org/10.1073/pnas.0506344102`.

J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Kaptein, B. D. Sykes, P. E. Wright, and K. Wüthrich. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J Mol Biol*, 280(5):933–952, Jul 1998. doi: 10.1006/jmbi.1998.1852.
URL `http://dx.doi.org/10.1006/jmbi.1998.1852`.

J. Moult, T. Hubbard, S. H. Bryant, K. Fidelis, and J. T. Pedersen. Critical assessment of methods of protein structure prediction (casp): round ii. *Proteins*, Suppl 1:2–6, 1997.

A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995. doi: 10.1006/jmbi.1995.0159.
URL `http://dx.doi.org/10.1006/jmbi.1995.0159`.

S. Neal, A. M. Nip, H. Zhang, and D. S. Wishart. Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *J Biomol NMR*, 26(3):215–240, Jul 2003.

S. Neal, M. Berjanskii, H. Zhang, and D. S. Wishart. Accurate prediction of protein torsion angles using chemical shifts and sequence homology. *Magn Reson Chem*, 44:S158–S167, Jul 2006. doi: 10.1002/mrc.1832.
URL `http://dx.doi.org/10.1002/mrc.1832`.

G. Nicastro, R. P. Menon, L. Masino, P. P. Knowles, N. Q. McDonald, and A. Pastore. The solution structure of the josephin domain of ataxin-3: structural determinants for molecular recognition. *Proc Natl Acad Sci*, 102(30):10493–10498, Jul 2005. doi: 10.1073/pnas.0501732102.
URL `http://dx.doi.org/10.1073/pnas.0501732102`.

E. Oldfield. Chemical shifts and three-dimensional protein structures. *J Biomol NMR*, 5(3):217–225, Apr 1995.

A. Pastore and A. M. Lesk. Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. *Proteins*, 8(2):133–155, 1990. doi: 10.1002/prot.340080204.
URL `http://dx.doi.org/10.1002/prot.340080204`.

B. Rost and V. Eyrich. EVA: large-scale analysis of secondary structure prediction. *Proteins*, 45(suppl 5):192–199, 2001.

D. Rumelhart, G. Hintont, and R. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

W. L. Ruzzo and M. Tompa. A linear time algorithm for finding all maximal scoring subsequences. *Proc Int Conf Intell Syst Mol Biol*, pages 234–241, 1999.

R. Sadreyev and N. Grishin. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, 326(1):317–336, Feb 2003.

C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore. The xplor-nih nmr molecular structure determination package. *J Magn Reson*, 160(1):65–73, Jan 2003.

B. Seavey, E. Farr, W. Westler, and J. Markley. A relational database for sequence-specific protein NMR data. *J Biomol NMR*, 1:217–236, 1991.

I. Shindyalov and P. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9): 739–747, 1998.

N. Siew, A. Elofsson, L. Rychlewski, and D. Fischer. Maxsub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9):776–785, Sep 2000.

J. Söding. Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21(7):951–960, Apr 2005. doi: 10.1093/bioinformatics/bti125. URL `http://dx.doi.org/10.1093/bioinformatics/bti125`.

G. Wagner, A. Pardi, and K. Wuethrich. Hydrogen bond length and proton NMR chemical shifts in proteins. *Journal of the American Chemical Society*, 105(18): 5948–5949, 1983.

G. Wang, Y. Jin, and R. L. Dunbrack. Assessment of fold recognition predictions in casp6. *Proteins*, 61 Suppl 7:46–66, 2005a. doi: 10.1002/prot.20721. URL `http://dx.doi.org/10.1002/prot.20721`.

L. Wang, H. R. Eghbalnia, A. Bahrami, and J. L. Markley. Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. *J Biomol NMR*, 32(1):13–22, May 2005b. doi: 10.1007/s10858-005-1717-0. URL `http://dx.doi.org/10.1007/s10858-005-1717-0`.

Y. Wang and O. Jardetzky. Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci*, 11(4):852–861, Apr 2002.

Y. Wang and D. S. Wishart. A simple method to adjust inconsistently referenced 13C and 15N chemical shift assignments of proteins. *J Biomol NMR*, 31(2): 143–148, Feb 2005. doi: 10.1007/s10858-004-7441-3. URL `http://dx.doi.org/10.1007/s10858-004-7441-3`.

D. S. Wishart and B. D. Sykes. The 13c chemical-shift index: a simple method for the identification of protein secondary structure using 13c chemical-shift data. *J Biomol NMR*, 4(2):171–180, Mar 1994.

D. S. Wishart, B. D. Sykes, and F. M. Richards. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J Mol Biol*, 222(2):311–333, Nov 1991.

D. S. Wishart, B. D. Sykes, and F. M. Richards. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6):1647–1651, Feb 1992.

D. S. Wishart, C. G. Bigam, A. Holm, R. S. Hodges, and B. D. Sykes. 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids I Investigations of nearest-neighbor effects. *J Biomol NMR*, 5(1):67–81, Jan 1995a.

D. S. Wishart, C. G. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. L. Markley, and B. D. Sykes. 1H, 13C and 15N chemical shift referencing in biomolecular NMR. *J Biomol NMR*, 6(2):135–140, Sep 1995b.

G. Yona and M. Levitt. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, 315(5):1257–1275, Feb 2002. doi: 10.1006/jmbi.2001.5293.
URL `http://dx.doi.org/10.1006/jmbi.2001.5293`.

H. Zhang, S. Neal, and D. S. Wishart. RefDB: a database of uniformly referenced protein chemical shifts. *J Biomol NMR*, 25(3):173–195, Mar 2003.

A SHIFTX Supplementary Material

Residue	Target	Partner
All	CA	N
All	HA	CA
All but PRO	H	N
ALA	HB1	CB
ALA	HB2	CB
ALA	HB3	CB
ARG	HB2	CB
ARG	HB3	CB
ARG	HD2	CD
ARG	HD3	CD
ARG	HE	NE
ARG	HG2	CG
ARG	HG3	CG
ARG	HH11	NH1
ARG	HH12	NH1
ARG	HH21	NH2
ARG	HH22	NH2
ASN	HB2	CB
ASN	HB3	CB
ASN	HD21	ND2
ASN	HD22	ND2
ASP	HB2	CB
ASP	HB3	CB
ASP	HD2	OD2
CYS	HB2	CB
CYS	HB3	CB
CYS	HG	SG
GLN	HB2	CB
GLN	HB3	CB

continued on next page

continued from previous page

Residue	Target	Partner
GLN	HE21	NE2
GLN	HE22	NE2
GLN	HG2	CG
GLN	HG3	CG
GLU	HB2	CB
GLU	HB3	CB
GLU	HE2	OE2
GLU	HG2	CG
GLU	HG3	CG
GLY	HA2	CA
GLY	HA3	CA
HIS	HB2	CB
HIS	HB3	CB
HIS	HD1	ND1
HIS	HD2	CD2
HIS	HE1	CE1
HIS	HE2	NE2
ILE	HB	CB
ILE	HD1	CD1
ILE	HD2	CD1
ILE	HD3	CD1
ILE	HG12	CG1
ILE	HG13	CG1
ILE	HG21	CG2
ILE	HG22	CG2
ILE	HG23	CG2
LEU	HB2	CB
LEU	HB3	CB
LEU	HD11	CD1
LEU	HD12	CD1
LEU	HD13	CD1
LEU	HD21	CD2
LEU	HD22	CD2
LEU	HD23	CD2
LEU	HG	CG
LYS	HB2	CB
LYS	HB3	CB

continued on next page

continued from previous page

Residue	Target	Partner
LYS	HD2	CD
LYS	HD3	CD
LYS	HE2	CE
LYS	HE3	CE
LYS	HG2	CG
LYS	HG3	CG
LYS	HZ1	NZ
LYS	HZ2	NZ
LYS	HZ3	NZ
MET	HB2	CB
MET	HB3	CB
MET	HE1	CE
MET	HE2	CE
MET	HE3	CE
MET	HG2	CG
MET	HG3	CG
PHE	HB2	CB
PHE	HB3	CB
PHE	HD1	CD1
PHE	HD2	CD2
PHE	HE1	CE1
PHE	HE2	CE2
PHE	HZ	CZ
PRO	HB2	CB
PRO	HB3	CB
PRO	HD2	CD
PRO	HD3	CD
PRO	HG2	CG
PRO	HG3	CG
SER	HB2	CB
SER	HB3	CB
SER	HG	OG
THR	HB	CB
THR	HG1	OG1
THR	HG21	CG2
THR	HG22	CG2
THR	HG23	CG2

continued on next page

continued from previous page

Residue	Target	Partner
TRP	HB2	CB
TRP	HB3	CB
TRP	HD1	CD1
TRP	HE1	NE1
TRP	HE3	CE3
TRP	HH2	C22
TRP	HZ2	CZ2
TRP	HZ3	CZ3
TYR	HB2	CB
TYR	HB3	CB
TYR	HD1	CD1
TYR	HD2	CD2
TYR	HE1	CE1
TYR	HE2	CE2
TYR	HH	OH
VAL	HB	CB
VAL	HG11	CG1
VAL	HG12	CG1
VAL	HG13	CG1
VAL	HG21	CG2
VAL	HG22	CG2
VAL	HG23	CG2

Table A.1: Target and partner atoms for electric field effects calculation.

B Empirical vs. Theoretical P-Values for Set S_1

Figure B.1: Alignments consisting of one block (3,654,077 data points)

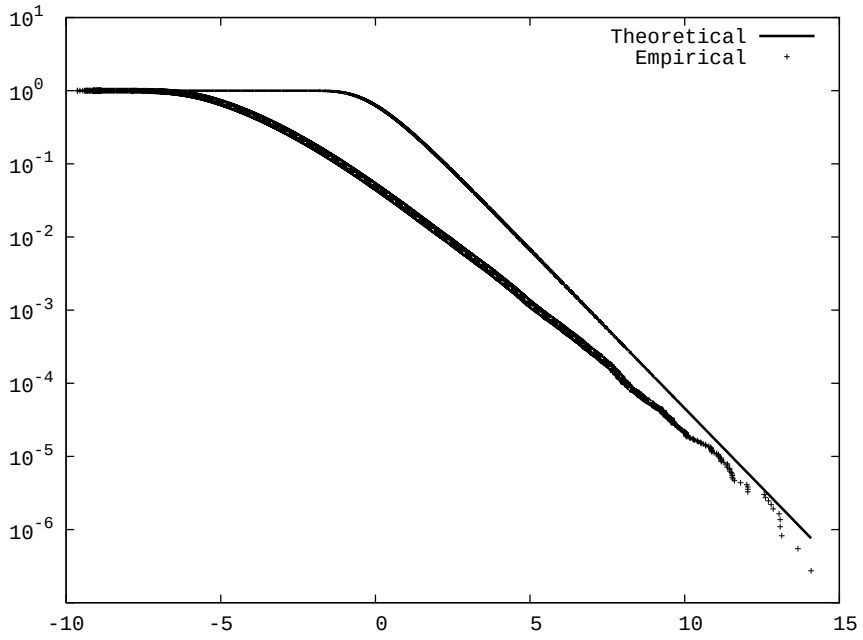


Figure B.2: Alignments consisting of two blocks (2,260,265 data points)

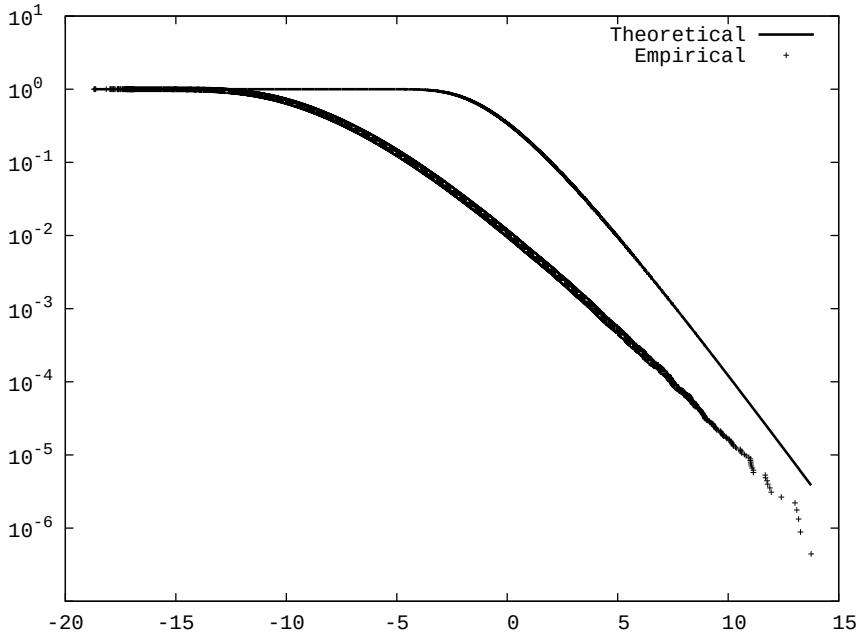


Figure B.3: Alignments consisting of three blocks (638,037 data points)

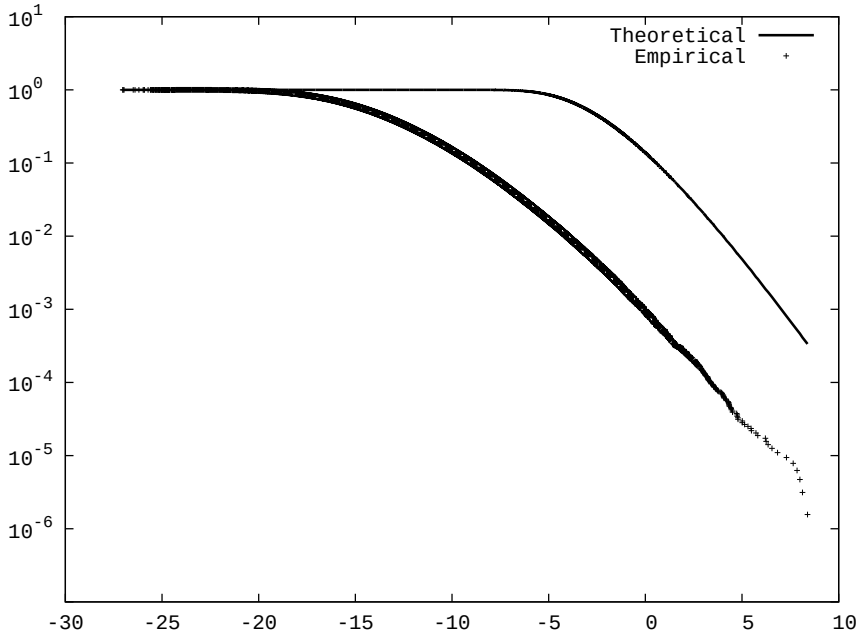


Figure B.4: Alignments consisting of four blocks (96,385 data points)

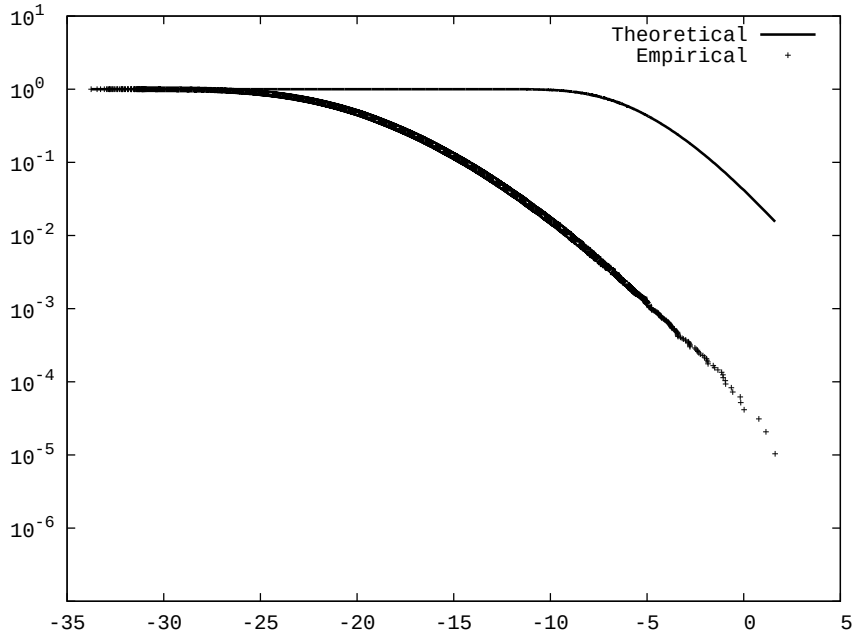


Figure B.5: Alignments consisting of five blocks (8,941 data points)

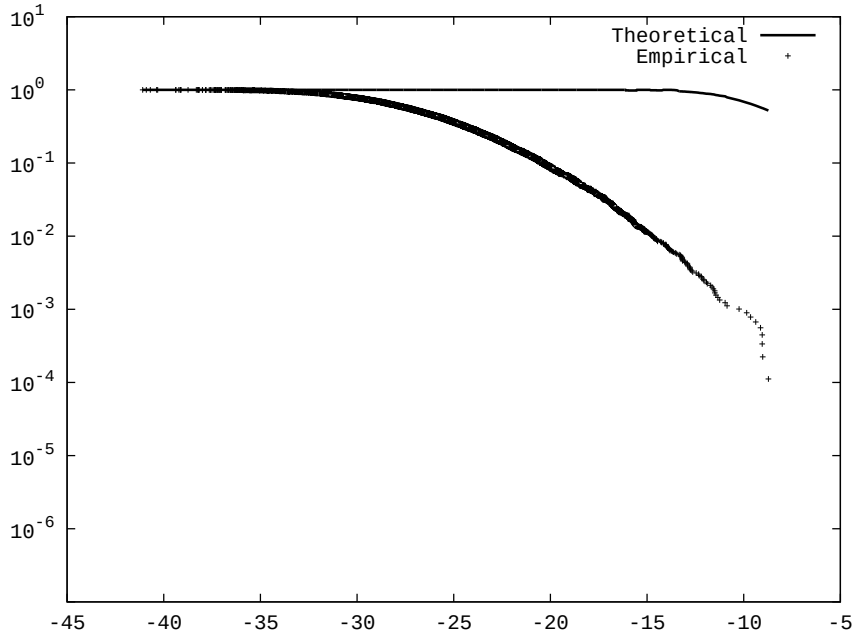


Figure B.6: Alignments consisting of six blocks (461 data points)

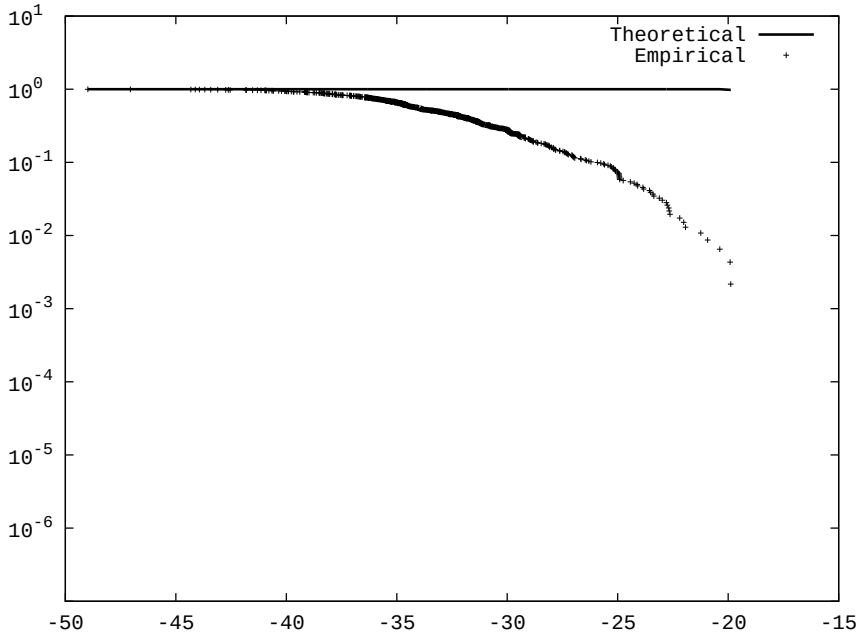
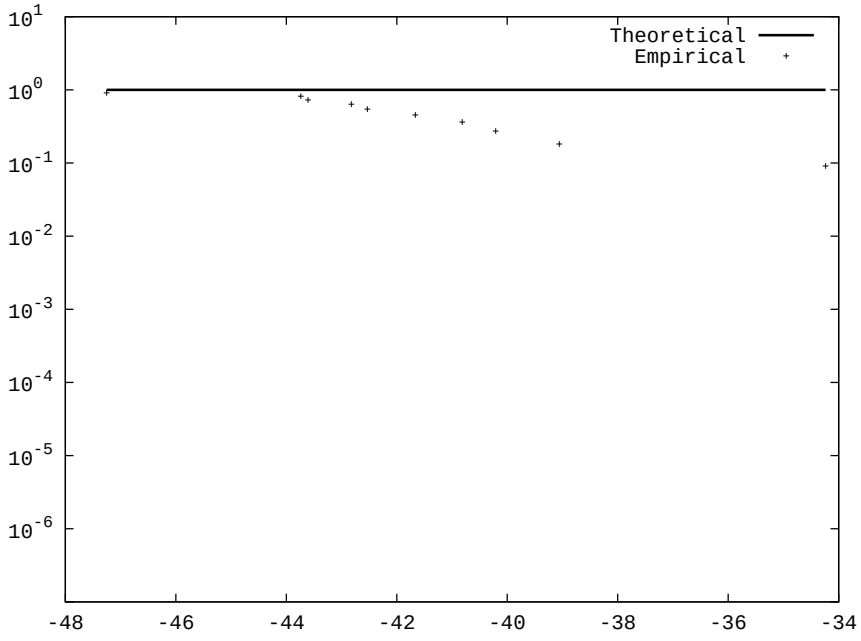


Figure B.7: Alignments consisting of seven blocks (10 data points)



C Empirical vs. Theoretical P-Values for Set S_2

Figure C.1: Alignments consisting of one block (3,654,077 data points)

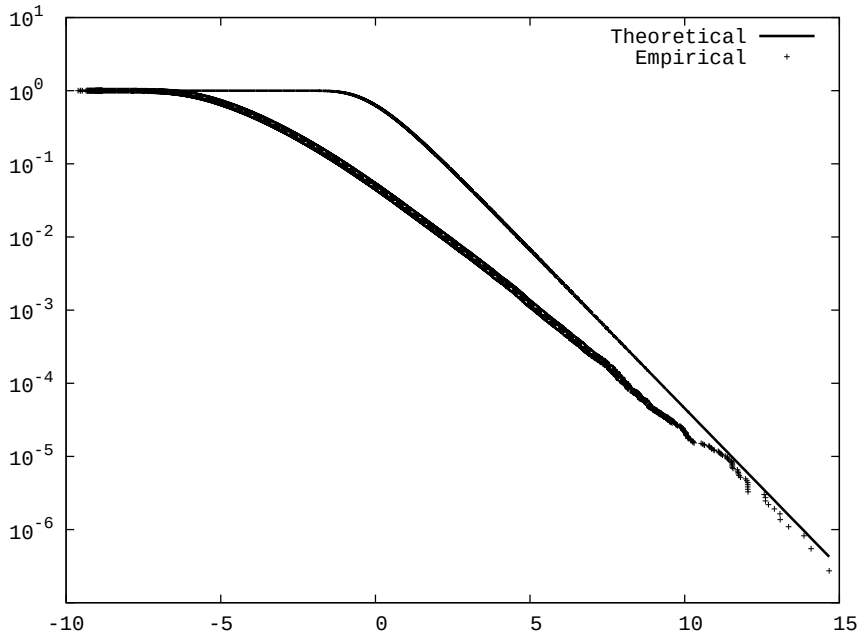


Figure C.2: Alignments consisting of two blocks (2,190,528 data points)

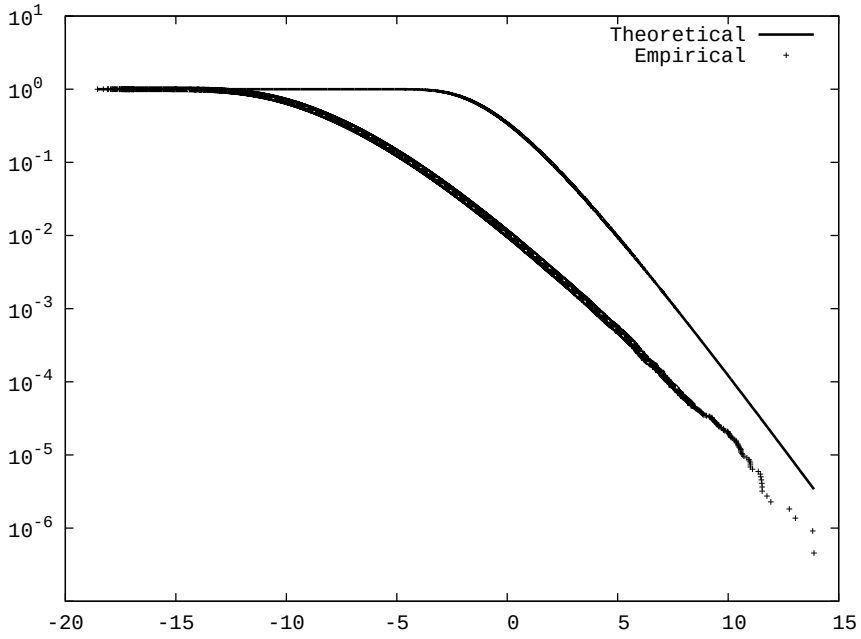


Figure C.3: Alignments consisting of tHree blocks (574,503 data points)

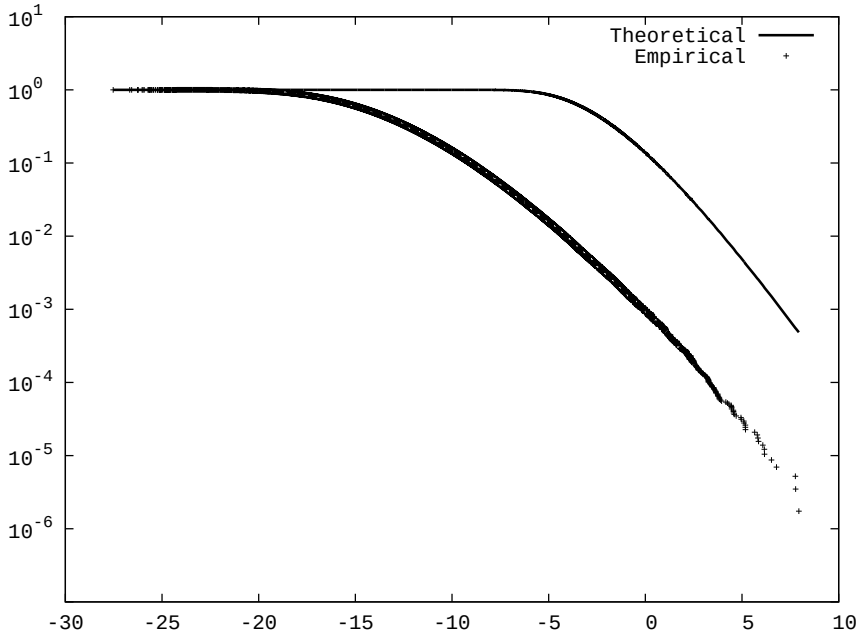


Figure C.4: Alignments consisting of four blocks (77,692 data points)

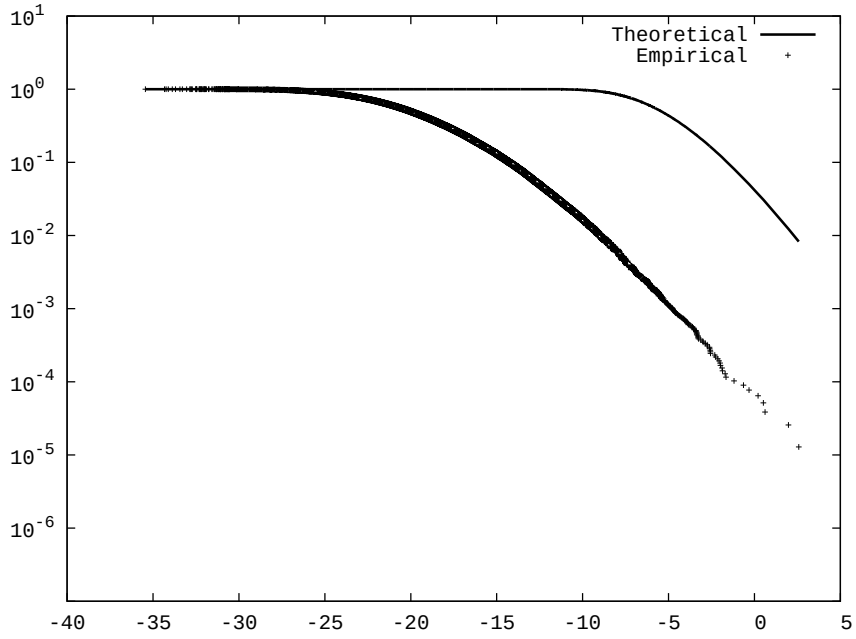


Figure C.5: Alignments consisting of five blocks (6,037 data points)

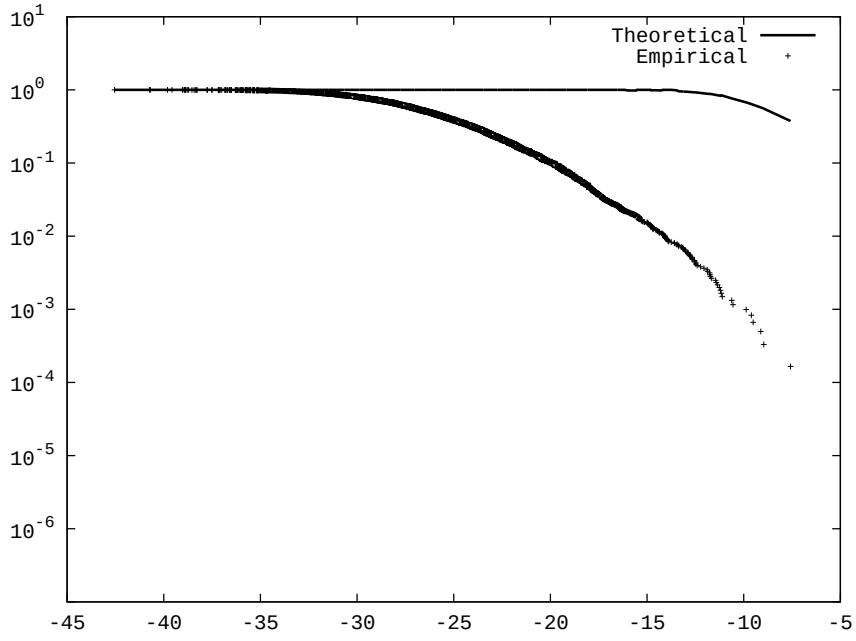


Figure C.6: Alignments consisting of six blocks (241 data points)

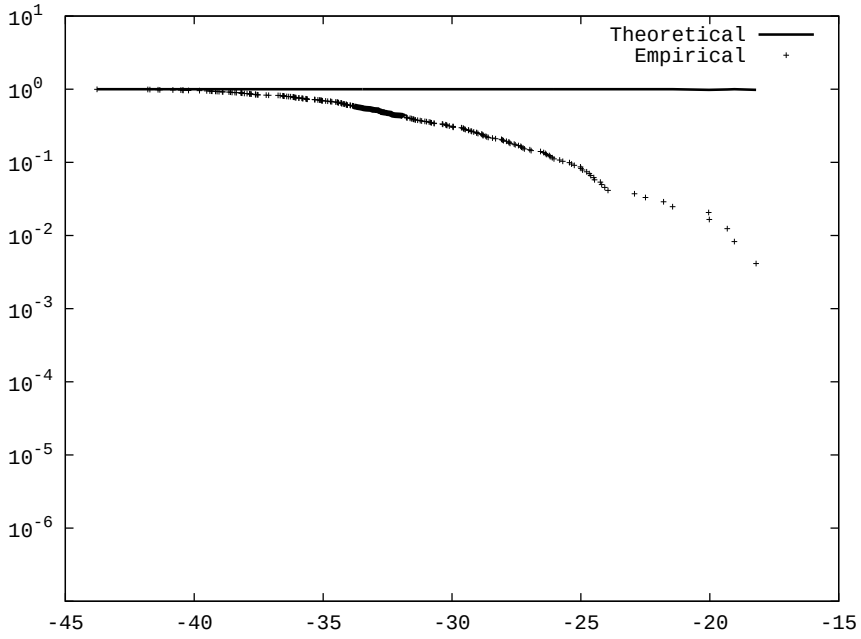
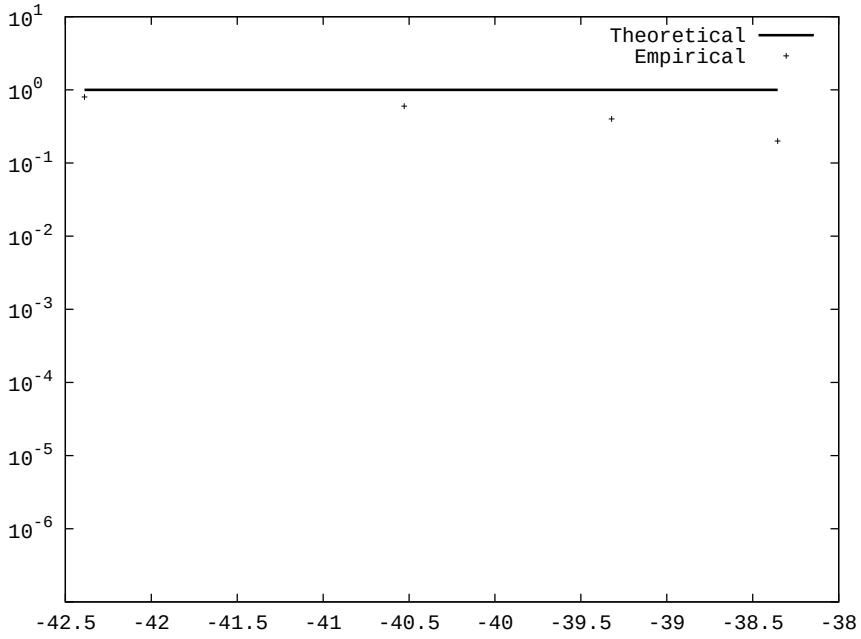


Figure C.7: Alignments consisting of seven blocks (4 data points)



D The Benchmark Set

BMRB Id.	Shift Set	ASTRAL Id.	%H	%S	%C	Length
274	1	d1wejf_	44	3	53	102
385	1	d7rsa__	22	35	43	122
434	1	d2i1b__	1	51	48	152
435	1	d2i1b__	1	52	47	151
443	1	d7rsa__	22	35	43	124
915	1	d1l58__	66	9	25	161
975	1	d1brsb_	23	22	55	110
1061	1	d2i1b__	1	51	48	153
1062	1	d2i1b__	1	51	48	152
1093	1	d4lzt__	46	9	45	125
1170	1	d1wejf_	44	4	52	100
1657	1	d2rn2__	37	32	31	155
1672	1	d256bb_	83	0	17	104
1766	1	d1noa__	0	57	43	113
2059	1	d1id2c_	2	50	48	105
2208	1	d1w8ma_	14	37	49	162
2542	1	d1jsf__	45	11	44	127
2868	1	d1irda_	78	0	22	105
3442	1	d1irda_	79	0	21	103
4022	1	d1czm__	14	32	54	259
4031	1	d7rsa__	22	35	43	124
4038	1	d1jl6a_	78	0	22	147
4056	1	d1lin__	60	5	35	141
4061	1	d1bzpa_	80	0	20	150
4062	1	d1bzpa_	79	0	21	147
4077	1	d1fkf__	13	42	45	107
4082	1	d1fil__	27	31	42	138
4083	1	d3chy__	45	17	38	127
4129	1	d2scpb_	67	3	30	170

continued on next page

continued from previous page

BMRB Id.	Shift Set	ASTRAL Id.	%H	%S	%C	Length
4132	1	d1u9aa_	36	21	43	157
4173	1	d1g49b_	24	19	57	169
4186	1	d1cbs__	13	58	29	131
4189	1	d1wejf_	43	3	54	104
4291	1	d2cthb_	24	9	67	107
4293	1	d1fsjb_	37	18	45	129
4299	1	d1q0na_	31	29	40	149
4300	1	d1q0na_	31	29	40	154
4331	1	d2sici_	16	33	51	104
4352	1	d1fsjb_	38	18	44	133
4364	1	d1g49b_	25	19	56	159
4411	1	d1tgj__	18	47	35	101
4421	1	d1gxqa_	39	24	37	103
4472	1	d3chy__	46	17	37	128
4554	1	d1ddrb_	24	35	41	159
4562	1	d4lzt__	45	9	46	129
4568	1	d1bzpa_	80	0	20	153
4573	1	d1dhn__	37	32	31	116
4580	2	d2dlfl_	2	52	46	112
4676	1	d1bzpa_	80	0	20	152
4679	1	d830cb_	26	18	56	154
4681	1	d1opbd_	12	58	30	133
4681	3	d1opbd_	12	59	29	132
4682	2	d1opbd_	12	58	30	133
4682	3	d1opbd_	12	59	29	132
4761	1	d1cpq__	75	1	24	129
4767	1	d256bb_	82	0	18	105
4837	1	d1gu2a_	47	6	47	124
4848	1	d1g6sa_	31	30	39	232
4854	1	d1g6sa_	31	30	39	233
4876	1	d1el1b_	40	9	51	130
4883	1	d1qqya_	41	8	51	126
4887	1	d1qqya_	41	8	51	126
4943	1	d4lzt__	46	9	45	128
4943	2	d4lzt__	46	9	45	128
4964	1	d1brsb_	23	22	55	110
4980	1	d1y92b_	25	34	41	123

continued on next page

continued from previous page

BMRB Id.	Shift Set	ASTRAL Id.	%H	%S	%C	Length
5011	1	d1rcf__	38	23	39	145
5026	1	d1wejf_	43	3	54	103
5064	1	d1gnua_	29	25	46	101
5068	1	d4lzt__	46	9	45	128
5069	1	d4lzt__	50	8	42	110
5083	1	d1b56__	15	61	24	133
5123	1	d1jsf__	47	9	44	123
5124	1	d1oua__	44	11	45	127
5125	1	d1jsf__	46	11	43	125
5128	1	d1gnua_	28	25	47	103
5130	1	d1jsf__	44	10	46	130
5142	1	d1jsf__	44	10	46	130
5169	1	d1i4fb_	0	49	51	100
5222	1	d1ezka_	37	23	40	142
5231	1	d1g49b_	26	19	55	153
5239	1	d2cthb_	24	9	67	105
5244	1	d2end__	54	2	44	137
5269	1	d1s69a_	75	0	25	123
5287	1	d1lin__	59	5	36	144
5333	1	d1j0oa_	27	11	62	107
5343	1	d1noa__	0	57	43	113
5344	1	d1noa__	0	60	40	105
5350	1	d5pnt__	45	16	39	147
5372	1	d1wejf_	43	3	54	104
5393	1	d1s0pb_	78	1	21	174
5404	1	d1g2ac_	30	34	36	146
5474	1	d1eena_	46	17	37	244
5497	1	d1uc7b_	42	17	41	122
5512	1	d1tw4b_	12	60	28	125
5540	1	d1bu5b_	39	25	36	147
5571	1	d1bu5b_	39	25	36	147
5578	1	d1crb__	11	55	34	134
5579	1	d1crb__	11	55	34	134
5625	1	d2cthb_	24	9	67	107
5679	1	d1v6wb1	11	42	47	128
5740	1	d1ra9__	25	34	41	156
5741	1	d1ra9__	26	34	40	158

continued on next page

continued from previous page

BMRB Id.	Shift Set	ASTRAL Id.	%H	%S	%C	Length
5759	1	d1dz4b_	45	15	40	175
5761	1	d1gyva_	2	59	39	115
5803	1	d4lzt__	47	8	45	118
5854	1	d1kqpb_	56	7	37	265
5856	1	d1irda_	81	0	19	138
5856	2	d1irdb_	78	0	22	145
5921	1	d1vc1b_	43	26	31	110
5969	1	d1noa__	0	57	43	113
5981	1	d1s3va_	20	35	45	186
6223	1	d1p6oa_	50	19	31	150
6230	1	d1irdb_	77	0	23	104
6230	3	d1irdb_	77	0	23	104
6232	1	d1vc1b_	44	27	29	104
6292	1	d1jiwi_	11	51	38	104
6313	1	d1dqeb_	69	1	30	128
6321	1	d1otba_	28	38	34	100
6357	1	d1xpb__	46	17	37	262
6444	1	d1rmza_	27	18	55	154
6494	1	d1dbfc_	33	25	42	127
6495	1	d1dbfc_	33	25	42	127
6496	1	d1dbfc_	33	27	40	118
6572	1	d1j0oa_	28	10	62	101
6622	1	d4lzt__	42	10	48	113
6642	1	d1tw4b_	12	59	29	121
6807	1	d3mbp__	46	21	33	363
6888	1	d1noa__	0	58	42	110
7003	1	d1byqa_	42	25	33	190
7107	1	d1b56__	15	62	23	132
7125	1	d1irda_	80	0	20	137
7125	2	d1irdb_	78	0	22	145
7126	1	d1brsb_	23	22	55	109
7133	1	d1i58a_	54	24	22	115
7234	1	d1o08a_	53	10	37	220
7235	1	d1o08a_	57	11	32	196
7293	1	d1omra_	66	4	30	177
7355	1	d1icm__	12	61	27	121
7356	1	d1icm__	12	60	28	131

continued on next page

continued from previous page

BMRB Id.	Shift Set	ASTRAL Id.	%H	%S	%C	Length
7357	1	d1icm__	11	61	28	129
15066	1	d1sw0b_	49	12	39	220
15067	1	d1sw0b_	47	13	40	206
15082	1	d1icm__	12	60	28	131

Table D.1: The Benchmark Set for the Chemical Shift Pipeline.

E Chemical Shift Pipeline Results

Table E.1: Result using only alignments with at most 30% sequence identity.

Set	Δ_Φ	Δ_Ψ	✔$_\Phi$	✔$_\Psi$	✘$_\Phi$	✘$_\Psi$	**C**
$\mathbf{2}_{100\%}$	17	22	73	70	11	13	12
$\mathbf{1}_{90\%}$	18	23	70	68	12	14	28
$\mathbf{1}_{80\%}$	20	25	69	66	14	16	33
$\mathbf{1}_{70\%}$	21	27	67	65	15	17	38
$\mathbf{2}_{90\%}$	22	29	66	64	16	19	42
$\mathbf{1}_{60\%}$	21	28	67	64	15	18	42
$\mathbf{2}_{80\%}$	22	30	65	63	17	19	46
$\mathbf{2}_{70\%}$	23	30	65	63	17	20	47
$\mathbf{1}_{50\%}$	23	30	65	63	17	19	44
$\mathbf{2}_{60\%}$	23	31	64	62	18	20	49
$\mathbf{1}_{40\%}$	23	31	65	62	17	20	45
$\mathbf{2}_{40\%}$	24	33	63	61	19	22	50
$\mathbf{2}_{50\%}$	24	32	64	61	18	21	50
$\mathbf{1}_{30\%}$	24	32	64	61	18	21	47
$\mathbf{3}_{100\%}$	25	33	63	60	19	22	51
$\mathbf{2}_{30\%}$	25	34	62	59	20	23	52
$\mathbf{1}_{100\%}$				n/a			

Table E.2: Result using only alignments with at most 40% sequence identity.

Set	Δ_Φ	Δ_Ψ	✔$_\Phi$	✔$_\Psi$	✘$_\Phi$	✘$_\Psi$	**C**
$\mathbf{2}_{100\%}$	16	21	71	70	10	12	27
$\mathbf{1}_{90\%}$	18	23	70	69	12	14	45
$\mathbf{1}_{80\%}$	19	25	69	68	13	15	51
$\mathbf{1}_{100\%}$	15	25	70	68	8	12	3
$\mathbf{1}_{60\%}$	20	27	68	67	14	16	53
$\mathbf{1}_{70\%}$	19	26	69	67	13	15	52
$\mathbf{2}_{90\%}$	20	27	68	66	14	16	54
$\mathbf{2}_{80\%}$	21	28	67	65	15	17	55
$\mathbf{1}_{50\%}$	21	29	67	65	15	18	54
$\mathbf{2}_{70\%}$	22	29	66	64	16	18	56
$\mathbf{2}_{60\%}$	22	30	66	64	16	19	56
$\mathbf{1}_{40\%}$	22	30	66	64	16	19	53
$\mathbf{3}_{100\%}$	23	32	65	63	17	20	57
$\mathbf{2}_{50\%}$	23	31	65	63	17	20	57
$\mathbf{2}_{40\%}$	23	32	65	63	17	20	57
$\mathbf{1}_{30\%}$	22	31	66	63	16	20	55
$\mathbf{2}_{30\%}$	24	33	64	61	18	21	58

Table E.3: Result using only alignments with at most 50% sequence identity.

Set	Δ_Φ	Δ_Ψ	✔$_\Phi$	✔$_\Psi$	✘$_\Phi$	✘$_\Psi$	**C**
$\mathbf{1}_{100\%}$	14	19	74	74	8	9	19
$\mathbf{2}_{100\%}$	17	21	72	72	10	11	41
$\mathbf{1}_{90\%}$	18	23	71	70	11	13	54
$\mathbf{1}_{80\%}$	18	24	70	69	12	14	57
$\mathbf{1}_{70\%}$	19	25	69	68	13	15	58
$\mathbf{1}_{60\%}$	19	26	69	68	13	15	59
$\mathbf{2}_{90\%}$	20	26	69	67	13	16	59
$\mathbf{2}_{70\%}$	21	28	67	66	15	17	60
$\mathbf{2}_{80\%}$	20	27	68	66	14	17	60
$\mathbf{1}_{50\%}$	20	27	68	66	14	17	60
$\mathbf{2}_{60\%}$	21	29	67	65	15	18	61
$\mathbf{1}_{40\%}$	21	28	67	65	15	18	59
$\mathbf{3}_{100\%}$	22	30	66	64	16	19	61
$\mathbf{2}_{50\%}$	22	30	66	64	16	19	61
$\mathbf{2}_{40\%}$	22	30	66	64	16	19	61
$\mathbf{1}_{30\%}$	22	29	67	64	15	19	60
$\mathbf{2}_{30\%}$	23	31	65	63	17	20	61

Table E.4: Result using only alignments with at most 60% sequence identity.

Set	Δ_Φ	Δ_Ψ	✔$_\Phi$	✔$_\Psi$	✘$_\Phi$	✘$_\Psi$	C
$\mathbf{1}_{100\%}$	14	18	76	76	7	8	34
$\mathbf{2}_{100\%}$	16	20	74	73	9	10	52
$\mathbf{1}_{90\%}$	17	22	73	72	10	12	60
$\mathbf{1}_{80\%}$	17	23	72	71	11	13	62
$\mathbf{1}_{60\%}$	18	24	71	70	12	14	63
$\mathbf{1}_{70\%}$	18	23	71	70	11	13	62
$\mathbf{2}_{90\%}$	18	24	70	69	12	14	63
$\mathbf{2}_{80\%}$	19	25	70	68	13	15	64
$\mathbf{2}_{70\%}$	20	26	69	68	13	15	64
$\mathbf{1}_{50\%}$	19	26	70	68	13	15	63
$\mathbf{1}_{40\%}$	19	26	69	68	13	16	63
$\mathbf{2}_{60\%}$	20	27	69	67	14	16	64
$\mathbf{2}_{50\%}$	20	28	68	67	14	17	64
$\mathbf{1}_{30\%}$	20	27	69	67	14	17	63
$\mathbf{3}_{100\%}$	21	28	68	66	15	17	65
$\mathbf{2}_{40\%}$	21	28	68	66	15	17	64
$\mathbf{2}_{30\%}$	21	29	67	65	15	18	65

Table E.5: Result using only alignments with at most 70% sequence identity.

Set	Δ_Φ	Δ_Ψ	✔$_\Phi$	✔$_\Psi$	✘$_\Phi$	✘$_\Psi$	**C**
$\mathbf{1}_{100\%}$	13	17	77	77	7	8	43
$\mathbf{2}_{100\%}$	15	19	75	75	8	10	55
$\mathbf{1}_{90\%}$	16	21	74	73	10	11	63
$\mathbf{1}_{80\%}$	17	22	73	72	10	12	64
$\mathbf{1}_{70\%}$	17	23	72	71	11	13	64
$\mathbf{1}_{60\%}$	18	23	72	71	11	13	65
$\mathbf{2}_{90\%}$	18	24	72	70	11	13	65
$\mathbf{2}_{70\%}$	19	25	70	69	13	15	66
$\mathbf{1}_{40\%}$	19	25	70	69	13	15	65
$\mathbf{2}_{80\%}$	19	25	71	69	12	14	65
$\mathbf{1}_{50\%}$	18	24	71	69	12	14	65
$\mathbf{2}_{50\%}$	20	27	69	68	14	16	66
$\mathbf{2}_{60\%}$	19	26	70	68	13	15	66
$\mathbf{1}_{30\%}$	19	26	70	68	13	16	65
$\mathbf{3}_{100\%}$	20	27	69	67	14	16	66
$\mathbf{2}_{40\%}$	20	27	69	67	14	17	66
$\mathbf{2}_{30\%}$	21	28	68	66	15	17	66

Table E.6: Result using only alignments with at most 80% sequence identity.

Set	Δ_Φ	Δ_Ψ	✔$_\Phi$	✔$_\Psi$	✘$_\Phi$	✘$_\Psi$	C
$\mathbf{1}_{100\%}$	13	17	77	77	7	8	46
$\mathbf{2}_{100\%}$	15	19	75	75	8	10	58
$\mathbf{1}_{90\%}$	16	21	74	73	10	11	64
$\mathbf{1}_{80\%}$	17	22	73	72	10	12	65
$\mathbf{1}_{70\%}$	17	22	72	71	11	13	66
$\mathbf{1}_{60\%}$	17	23	72	71	11	13	66
$\mathbf{2}_{90\%}$	18	23	72	70	11	13	66
$\mathbf{2}_{70\%}$	19	25	71	69	13	15	67
$\mathbf{1}_{40\%}$	19	25	70	69	13	15	66
$\mathbf{2}_{80\%}$	18	24	71	69	12	14	67
$\mathbf{1}_{50\%}$	18	24	71	69	12	14	66
$\mathbf{2}_{60\%}$	19	26	70	68	13	15	67
$\mathbf{2}_{50\%}$	20	26	70	68	13	16	67
$\mathbf{1}_{30\%}$	19	26	70	68	13	15	66
$\mathbf{3}_{100\%}$	20	27	69	67	14	16	67
$\mathbf{2}_{40\%}$	20	27	69	67	14	16	67
$\mathbf{2}_{30\%}$	20	28	69	67	14	17	67

Table E.7: Result using only alignments with at most 90% sequence identity.

Set	Δ_Φ	Δ_Ψ	✔$_\Phi$	✔$_\Psi$	✘$_\Phi$	✘$_\Psi$	**C**
$\mathbf{1}_{100\%}$	13	17	78	78	6	8	51
$\mathbf{2}_{100\%}$	14	19	76	75	8	9	60
$\mathbf{1}_{90\%}$	16	20	75	74	9	11	65
$\mathbf{1}_{80\%}$	16	21	74	73	10	11	66
$\mathbf{1}_{60\%}$	17	22	73	72	11	12	67
$\mathbf{1}_{70\%}$	17	22	73	72	10	12	66
$\mathbf{2}_{90\%}$	17	23	72	71	11	13	67
$\mathbf{1}_{50\%}$	18	24	72	71	11	13	67
$\mathbf{2}_{80\%}$	18	24	72	70	12	13	67
$\mathbf{2}_{70\%}$	18	24	71	70	12	14	68
$\mathbf{2}_{60\%}$	19	25	71	70	12	14	68
$\mathbf{1}_{40\%}$	18	24	71	70	12	14	67
$\mathbf{2}_{50\%}$	19	26	71	69	13	15	68
$\mathbf{1}_{30\%}$	18	25	71	69	12	15	67
$\mathbf{2}_{30\%}$	20	27	69	68	14	16	68
$\mathbf{3}_{100\%}$	19	26	70	68	13	15	68
$\mathbf{2}_{40\%}$	19	26	70	68	13	16	68

Table E.8: Result using all alignments.

Set	Δ_Φ	Δ_Ψ	✔$_\Phi$	✔$_\Psi$	✘$_\Phi$	✘$_\Psi$	C
$\mathbf{1}_{100\%}$	11	14	82	82	5	6	77
$\mathbf{2}_{100\%}$	12	16	80	80	7	8	77
$\mathbf{1}_{90\%}$	14	17	79	78	8	9	76
$\mathbf{1}_{80\%}$	14	18	78	77	8	10	76
$\mathbf{1}_{70\%}$	14	19	77	76	9	10	75
$\mathbf{1}_{60\%}$	15	19	77	76	9	10	75
$\mathbf{2}_{80\%}$	16	20	76	75	10	11	75
$\mathbf{1}_{50\%}$	15	20	76	75	10	11	74
$\mathbf{2}_{90\%}$	15	20	76	75	9	11	75
$\mathbf{2}_{70\%}$	16	21	75	74	10	12	74
$\mathbf{2}_{60\%}$	16	21	75	74	10	12	73
$\mathbf{1}_{40\%}$	16	21	75	74	10	12	72
$\mathbf{1}_{30\%}$	16	21	75	74	10	12	71
$\mathbf{3}_{100\%}$	17	23	74	73	11	13	73
$\mathbf{2}_{50\%}$	17	22	75	73	11	13	73
$\mathbf{2}_{40\%}$	17	22	74	73	11	13	73
$\mathbf{2}_{30\%}$	17	23	74	72	11	14	72

Lebenslauf

Name	Simon Wolfgang Ginzinger
Geboren	29.03.1978
Geburtsort	Rosenheim, Deutschland
Schule	
1990-1996	Bischöfliches Gymnasium Paulinum, Schwaz, Tirol, Österreich
20.06.1996	Matura (Abitur)
Universitäre Ausbildung	
1996-2002	Studium der Angewandten Informatik, Anwendungsfach Bioinformatik, Universität Salzburg
1999-2000	Studium an der Bowling Green State University, Ohio, USA
12.08.2000	Master of Science in Computer Science, Bowling Green State University
25.11.2002	Diplomprüfung Universität Salzburg
seit Okt. 2003	Wissenschaftlicher Angestellter an der Ludwig-Maximilians-Universität München, Promotion im Fach Bioinformatik
Sep. 2007	Visiting Scientist an der University of Alberta, Forschungsgruppe von Prof. Dr. David Wishart
Berufserfahrung	
Nov.2001-Apr.2001	Praktikum bei Proceryon Biosciences, Salzburg
Nov.2001-Feb.2003	Software Entwickler bei Proceryon Biosciences, Salzburg
Feb.2003-Aug.2003	Software Entwickler bei ARS (Adaptive Regelsysteme), Salzburg