Proteins are the molecular machines of the cell, carrying out many diverse functions. A protein’s particular function usually follows from its 3-dimensional structure, a relationship known as the structure-function paradigm.

A protein is a folded chain of amino acid residues connected by peptide bonds. Its structure has a regular backbone and varying amino acid side chains. There are four levels of protein structure: primary, secondary, tertiary and quaternary structure. Primary structure is the sequence of amino acid residues. Secondary structure comprises regular elements such as α-helices and β-sheets. Tertiary structure is determined by the precise location of every atom in the structure. Finally, quaternary structure is an assembly of several amino acid chains.

Structural similarities between proteins are extremely common. This is for various reasons. First, there is only a limited number of overall structural arrangements, or so-called protein folds. Then, parts of proteins which constitute evolutionary entities, so-called domains, re-appear in many different proteins, because they are deleted, inserted, swapped, mutated, etc. Further, there is a high evolutionary pressure on conserving protein function and hence on conserving protein structure during evolution. Proteins which share a common ancestor, i.e., homologs, thus often have a similar structure. Finally, convergent evolution may lead to similar structures for similar functions.

Because of this structure-function relationship, protein similarities help us to learn about the evolution and biological tasks of proteins. Therefore we would like to reliably detect such structural similarities. For this, we resort to experimental data which provides the 3D location of every atom in the protein structure. Two proteins are considered structurally similar if they have a similar protein backbone conformation. We thus select only one representative atom of every amino acid residue. The resulting two chains of representative atoms are then compared. From such a comparison we obtain a sequential one-to-one mapping between structurally equivalent residues in the two proteins. This mapping is called a structure alignment. It can be used to quantify the similarity of two proteins by assigning a corresponding similarity score to it. Given an alignment we can superpose two structures in 3-dimensional space in order to obtain a visual impression of their similarity.

Using optimization, we aim to detect the best structure alignment. Two steps are important. First, we need to define a scoring scheme according to which biologically correct alignments are top-scoring. Second, we use an algorithm that finds the score-optimal alignment. For reasonable scoring schemes, finding the optimal struc-
ture alignment is difficult; it is often assumed to be an NP-hard problem. As a result, almost all of the many existing structure alignment algorithms are heuristics. Even worse, each heuristic uses its own scoring scheme that it optimizes. Currently, there is no consensus which algorithm or scoring scheme is best.

Our contribution to the structure alignment problem is two-fold. First, we cast the problem in mathematical models and design exact algorithms to solve it. To this end we formulate general integer linear programs for inter-residue distance matrix-based structure alignment. These models cast many existing structure alignment approaches into a common framework. Finally, we solve these models to optimality by designing exact algorithms. An inter-residue distance matrix alignment is a sequential assignment of a subset of distance matrix rows and columns from one matrix to a subset of distance matrix rows and columns from the other matrix. This assignment should maximize the overall score for paired inter-residue distances. An exact algorithm for this problem will either return a mapping of maximum score or, if not found within time limit, bounds on this maximum score. We apply techniques from combinatorial optimization for our exact algorithms: integer linear programming, Lagrangian relaxation, branch-and-bound and branch-and-cut.

Our second contribution is the application of our algorithms to problems that can only be tackled by the use of exact algorithms. For example, we compute provably better alignments, obtain alignment quality guarantees and accurately quantify protein similarities. Further, we evaluate heuristic algorithms and rigorously compare scoring schemes for protein structure alignment. Finally, we provide all these services to structural biologists via a web server.

In summary, the work described in this thesis entitled *Exact algorithms for pairwise protein structure alignment* contributes towards improving algorithms and scoring schemes for protein structure alignment.