Chapter 1

Introduction

1.1 Problem Formulation

The alignment of two objects is the adjustment of the relative orientation of one of the objects to the other one with respect to a given similarity measure. Alignment of objects is a concept widely used in many scientific areas, whereby the dimensionality of the objects varies from 1-dimensional objects, such as alpha-numerical strings, over 2-dimensional objects, such as images, to 3-dimensional or even higher dimensional objects. Even though the method used for aligning objects strongly depends on their dimensionality and the application in focus, there are some general statements that can be made about alignment methods:

- *Similarity.* Alignment methods are used to determine the similarity of objects. Every alignment method is based on some similarity measure, which strongly affects the design of the method. Alignment methods should detect those relative positions for which the value of the considered similarity measure is large. While the primary aim of an alignment is to reveal the similarity of objects, their dissimilarity also becomes more apparent when the objects are aligned.

- *Local versus global alignment.* Alignment methods can be classified as either being local or global. Global methods take the whole object into account and, generally, try to identify the maximal *overall* similarity. Local methods, in contrast, align objects according to *local* similarity. Whether an alignment method is local or global is determined by the similarity measure, but the transition between global and local methods is smooth.

- *Correspondence.* A concept closely related to the problem of aligning objects is that of establishing a correspondence between parts of the objects. While global alignment methods always aim at establishing a correspondence in which all parts of the objects are represented, local alignment algorithms identify partial correspondences between the objects. That is to say, local methods will not necessarily find a correspondence between all parts of the objects, but concentrate on identifying correspondences with high similarity between the corresponding parts.

- *Pairwise versus multiple alignment.* In many cases, it is sufficient to be able to align two objects to each other. This process is called pairwise alignment. If more than two objects should be related to each other, more than a series of pairwise alignments is needed. In some way, all objects need to be related to each other at the same time. Multiple alignment aims at identifying commonalities among a set of objects. Even though these commonalities may be rather faint – depending on the number of objects and their diversity – these faint commonalities might reveal more information than similarities based on pairwise alignments, since they represent the properties common to all objects. This is of particular value if we want to relate the objects' properties to some common behavior.

This thesis focuses on alignment algorithms in virtual drug design, a subfield of biochemistry, which aims at designing new drug molecules "in silico", i.e. with the help of computers. In particular, we focus on alignment algorithms that might help in elucidating so called pharmacophores. The term *pharmacophore* was coined in the early 1900s by Paul Ehrlich. He defined a pharmacophore as *"a molecular framework that carries* (phoros) *the essential features responsible for a drug's* (pharmacon) *biological activity"* [54]. This definition was not modified until in 1977 Peter Gund specified it by describing a pharmacophore as *"a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity"* [69]. Pharmacophores are generally used as templates for screening large data bases of molecules to filter out those molecules bearing some specific structural and physico-chemical features. These molecules can then be investigated more thoroughly for their potential of replacing a previously known drug.

Pharmacophore identification can be divided into two fields: receptor-based identification and analog-based identification. While the first requires knowledge of the receptor, which is not always available, the latter requires a set of active analogs. Hence, this approach is also known as the *active analog approach* [70]. The active analog approach is based on the *molecular similarity principle*, which *"states that structurally similar molecules tend to have similar properties – physicochemical as well as biological ones – more often than structurally dissimilar molecules"* [21]. However, to be able to identify the essential features and to exclude features that merely fit by chance, we need a set of "diverse" active analogs which yet bind to the receptor via similar binding modes. The more active analogs we have and the more diverse they are, the more precisely the pharmacophore can be modeled. But the identification of common structural as well as physico-chemical features among a set of molecules can only be achieved by a multiple alignment.

If we want to design an alignment algorithm well-suited for pharmacophore identification, what properties should such an algorithm have? One requirement has already been stated: the algorithm must be able to align a set of molecules to each other, hence, we need a multiple alignment algorithm. But what additional properties are needed? Should the algorithm be local or global? Which similarity measure should be used, and which representation should the similarity measure be based on?

It has long been known that the functionality of molecules is strongly influenced by the molecule's shape. In 1890, Emil Fischer used the *lock-and-key* metaphor to explain

the binding of a small molecule, the ligand, to a larger molecule, the receptor. The ligand must fit into the active site of the receptor as a key must fit into its lock. Even though it has been known for quite some time that the lock often changes its shape during the binding process, known as *induced fit*, it nevertheless remains true, that receptor and ligand represent *partially* complementary shapes in the bound state. Van Drie argues that it is necessary to better account for the shape of the molecules in the search for a pharmacophore: *"if a molecule has an atom bumping into the wall of the receptor, no amount of fiddling with the features in the pharmacophore can account for this."* [49] The complementarity of molecular shapes is best expressed using the concept of molecular surfaces. Yet, geometric complementarity of the shapes alone might not suffice. An algorithm considering only the shape of the molecule is most likely to fail in producing good alignments for pharmacophore identification. Physico-chemical properties of the molecules need to match as well. Hence, we need a molecular *surface* alignment algorithm capable of incorporating physico-chemical properties.

As the key sticks out of the lock, the ligand is not completely enclosed by the active site. Those parts of the ligand sticking out of the receptor's active site can look very differently. What matters is the part interacting with the receptor. This leads us to a further requirement for the design of an alignment algorithm suitable for pharmacophore elucidation: it must be a *local* alignment algorithm able to establish *partial* correspondences between objects.

In this thesis we develop an alignment algorithm that addresses the requirements for pharmacophore elucidation mentioned above. To summarize, such an algorithm must be based on local similarity of the "outer shape" of the molecule, namely its surface. Furthermore, it must be able to consider geometrical as well as physico-chemical properties of the molecules, since both are important for binding. Finally, a multiple alignment algorithm is needed in order to eliminate features that are not essential in the binding process.

1.2 Example: Odor Molecules

In order to understand how odors are coded within the olfactory system of some organism, knowledge about the input to this system is needed. The input to the olfactory system can be described by the odors activating or deactivating the olfactory receptor neurons expressing the same olfactory receptor. This set of activating and deactivating odors is termed the *molecular receptive range*. In her doctoral thesis [137], Daniela Pelz studied a large number of odors (more than 100) from a variety of chemical classes in order to describe the molecular receptive range of the *olfactory receptor neurons* expressing the olfactory receptor Or22a in the model organism *Drosophila melanogaster* (fruit fly). Among the investigated odors she was able to identify 39 odors activating Or22a, which means that Or22a has a broad yet selective range. In addition to identifying the activating odors, Daniela Pelz also determined the dose necessary to stimulate a response of the respective olfactory receptor neurons. Thus, she was able to classify the odors according to this dose and found that ethyl- and methyl-hexanoate are the most potent odors.

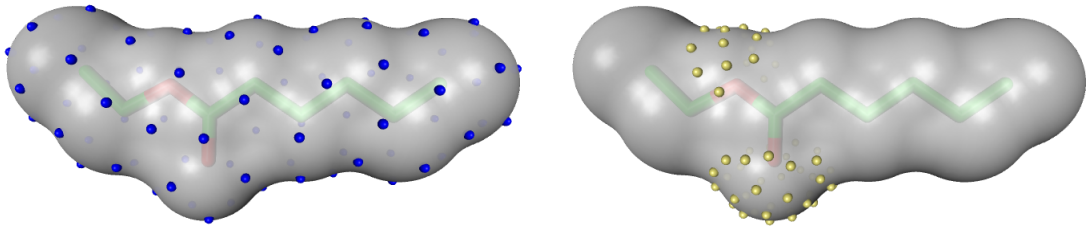

Figure 1.1: Representation of molecular surface and physico-chemical properties for ethyl-hexanoate. *Left:* Molecular surface with shape points. *Right:* Points representing potential hydrogen bond donor regions.

At this point the question arose, what structural and physico-chemical properties an odor needs to have to activate the receptor Or22a, whose 3-dimensional structure is unknown. This was the beginning of a very interesting collaboration and the starting point to dive into the field of structural alignment. The application of the multiple surface alignment algorithm proposed in this thesis allows for more insight into the molecular receptive range of Or22a, although it is difficult to judge the correctness of the identified model. In the following paragraph, we will shortly describe the results of the multiple alignment of all odors activating the olfactory receptor Or22a.

Since the 3-dimensional structure of Or22a is, as yet, unknown, we did not have any information about the 3-dimensional structures of the odors in the bound state. Moreover, the molecular structures of the 39 odors activating Or22a are, in large parts, very flexible; this means, they can adopt many different shapes. The first step we therefore had to accomplish was to compute those forms the odors might adopt when binding to the receptor. With the program ConFlow [125], we generated so called conformers for all odors, which were in the range from 7 conformers for 2-3-butanedione up to 491 conformers for butyl-butyrate. Since ethyl-hexanoate was found to be the most potent odor [137], we used ethyl-hexanoate as reference structure and compared the conformers of all other odors with the 154 conformers of ethyl-hexanoate. Apart from the shape of the molecules, we also considered the ability of the molecules to form hydrogen bonds. Surface shape as well as physico-chemical properties were represented by homogeneously distributed points (cf. Figure 1.1) to which we applied surface point matching. From the pairwise alignments of all odor molecules with ethyl-hexanoate, generated by the point matching algorithm applied to the surface points, we computed multiple alignments. Since some of the odor molecules can occur in different enantiomers, the number of molecules to be compared increased from 39 to 43.

The best multiple alignment of all odor molecules activating the olfactory receptor Or22a is shown in Figure 1.2 (third and fourth row). The odor molecules are shown by the ball-and-stick representation of their molecular structures, whereby the hydrogen atoms have been omitted for clarity. The upper two rows in Figure 1.2 show the same multiple alignment as in the bottom rows but restricted to the 20 molecules constituting the largest sub-matching. Remarkably, most of these 20 odor molecules belong to the

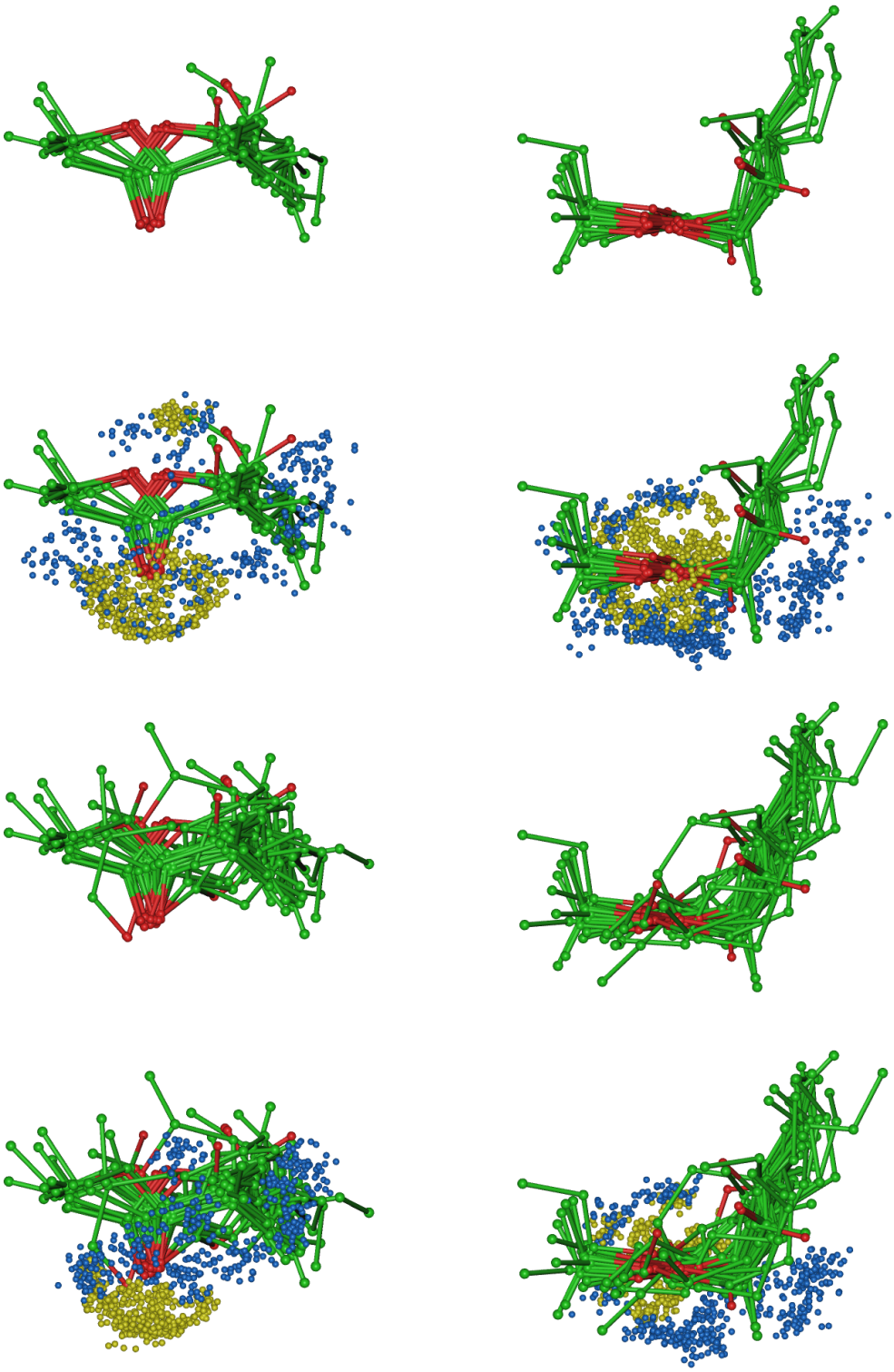

Figure 1.2: Multiple alignment of 20 odor molecules (first and second row) and 43 odor molecules (third and fourth row), respectively. The blue dots denote commonalities in shape, while the yellow dots denote commonalities in the ability to form hydrogen bonds. The right column shows the alignments viewed from the top w.r.t. the left column.

odors needing the lowest dose to activate the receptor. This gives a clear indication that the generated alignment is sensible in terms of receptor activation. The 20 molecules have in common that they allow to form two hydrogen bonds depicted by the yellow dots (second row, left image). After adding the other 23 molecules to the multiple alignment, only one hydrogen bonding region remains.

The alignment of the molecular surfaces of all 43 odor molecules, including a few cyclic molecular structures, shows a very conserved form. Yet, there are two molecules, namely E-4-methyl-cylcohexanol and gamma-R-valerolactone, which disturb the image slightly. However, both molecules do not belong to the highly active odor molecules.

As mentioned before, it is difficult to judge the quality of the multiple alignment depicted in Figure 1.2. In order to do so we would have to correlate the dose-curves with the structural and physico-chemical similarities and dissimilarities identified by the alignment. This could be done by applying a *quantitative structure activity relationship* (QSAR) analysis (see, e.g., [107]). However, this was outside the scope of this thesis, in which we focus on the generation of feasible alignments being a prerequisite for further analyses.

1.3 Outline of the Thesis

This thesis consists of five major parts, excluding the introductory chapter. While the first two parts describe the background as well as the previously published work in the field of molecular similarity, the latter three parts present the algorithmic details of our surface alignment approach.

Chapter 2, *Molecular Analysis and Visualization*, begins with an overview of molecular representations, which can be employed for visualization as well as for molecular analysis and similarity searching. In the second section, we present some physico-chemical interactions, in particular those, that are important for molecular docking and hence also for pharmacophore elucidation. This section is succeeded by an introduction to visualization methods that are used throughout this thesis. Here, we also describe the basis for our molecular alignment approach, i.e. the solvent excluded surface. Since we deal with flexible molecules, different forms of the same molecule need to be considered. How these forms can be computed will be the subject of the last section. Here, we also describe the concept of metastable molecular conformations and techniques for their visualization.

In Chapter 3, *Molecular Similarity*, we present previous work in the field of molecular similarity. Methods based on 2-dimensional as well as 3-dimensional representations are described, whereby we focus on methods applicable to drug-sized molecules. Using these representations, two major approaches can be followed. First, the generation of molecular descriptors and their comparison, and second, molecular alignment. We try to give a broad overview of both kinds of approaches. Alignment techniques utilizing molecular surfaces, however, are covered in more depth than other methods. The chapter is concluded by a summary and a short discussion about the applicability of the mentioned methods to certain classes of problems.

Chapter 4, *Point-Based Surface Representation*, describes the generation of a discrete

representation of the molecular surface using points. We present a new approach to distribute points regularly on the molecular surface w.r.t. to a scalar field given on the molecular surface. This approach consists of two steps, an initial point distribution step and a point relaxation step. For the initial point distribution we propose two different methods. While the first employs a Monte Carlo method, the second uses an efficient graph partitioning scheme. For the point relaxation step, we utilize centroidal Voronoi tesselation (CVT), which we extend to 2-manifold triangular meshes. We first recall some theoretic results on CVT, which are followed by the algorithmic details of how CVT can be applied to triangular meshes. In the fourth section we present experimental results, and we conclude this chapter with a short discussion.

In Chapter 5, *Pairwise Surface Alignment*, a new approach to molecular surface alignment based on the surface representation described in Chapter 4 is presented. This two-step approach first generates various initial alignments, each of which is locally optimized in the second step. The generation of initial alignments can be done atom based as well as surface based. Both methods are described and compared in the results section. For the optimization step, a point matching scheme is applied which we extend to surface points. We also introduce an efficient data structure which leads to a three-fold speed up. The quality of the proposed alignment algorithm is evaluated on two sets of molecules, a set of thermolysin inhibitors and a set of HIV-1 protease inhibitors, both of which are described in detail. In the results section, the algorithmic results are presented, which are discussed in the final section of this chapter.

Finally, in Chapter 6, *Multiple Surface Alignment*, we show how multiple alignments can be computed from pairwise alignments. Here, we employ the pairwise matchings which are successively intersected. We use an efficient data structure, called PATRICIA tree, which supports the process of intersecting pairwise matchings. Furthermore, the concept of Pareto set decomposition is described, which we utilize to sort the possibly large number of multiple matchings. For the evaluation of the algorithm, the same sets of molecules as in the previous chapter are used. The results are described in detail, and a discussion of the results concludes the chapter.

Key ideas of this thesis have been published and presented at the *1st and 2nd International Symposium on Computational Life Science*.

The first paper, "Multiple Semi-flexible 3D Superposition of Drug-sized Molecules" [15], describes the general idea of multiple alignment based on point matchings. Here, however, points represent atoms. Thus, the number of points is much smaller than if surface points are used.

The second paper, "A Point-Matching Based Algorithm for 3D Surface Alignment of Drug-Sized Molecules" [16], describes the extension of point matching to surface points, including the problem of distributing points regularly on a molecular surface. In this paper, we only describe the point distribution process for constant scalar fields.

New and unpublished results presented in this thesis include the generation of point distributions according to arbitrary scalar fields, and the extension of multiple atom alignment to multiple surface alignment, which is much more demanding due to the larger number of points that need to be handled.