Algorithms **2009**, *2*, 692-709; doi:10.3390/a2020692

OPEN ACCESS

algorithms

ISSN 1999-4893

www.mdpi.com/journal/algorithms

Article

Fast Structural Alignment of Biomolecules Using a Hash Table, N-Grams and String Descriptors

Raphael André Bauer [1,2,⋆], **Kristian Rother** [3,4,⋆], **Peter Moor** [1], **Knut Reinert** [5], **Thomas Steinke** [6], **Janusz M. Bujnicki** [3,4] **and Robert Preissner** [1,⋆]

[1] Charité Medical University, Structural Bioinformatics Group, Arnimallee 22, 14195 Berlin, Germany

[2] Graduate School: Genomics and Systems Biology of Molecular Networks, Invalidenstrasse 43, 10115 Berlin, Germany

[3] International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland

[4] Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

[5] Freie Universität Berlin, Algorithmische Bioinformatik, Institut für Informatik, Takustr. 9, 14195 Berlin, Germany

[6] Zuse Institute Berlin, Dept. Computer Science, Takustrasse 7, 14195 Berlin, Germany

⋆ Author to whom correspondence should be addressed; E-mails: ra.bauer@fu-berlin.de; krother@genesilico.pl; robert.preissner@charite.de

Received: 30 November 2008; in revised form: 8 April 2009 / Accepted: 9 April 2009 / Published: 21 April 2009

Abstract: This work presents a generalized approach for the fast structural alignment of thousands of macromolecular structures. The method uses string representations of a macromolecular structure and a hash table that stores n-grams of a certain size for searching. To this end, macromolecular structure-to-string translators were implemented for protein and RNA structures. A query against the index is performed in two hierarchical steps to unite speed and precision. In the first step the query structure is translated into n-grams, and all target structures containing these n-grams are retrieved from the hash table. In the second step all corresponding n-grams of the query and each target structure are subsequently aligned, and after each alignment a score is calculated based on the matching n-grams of query and target. The extendable framework enables the user to query and structurally align thousands of protein and RNA structures on a commodity machine and is available as open source from http://lajolla.sf.net.

Keywords: Structural alignment; protein; RNA; hash table; n-gram; torsion angles.

1. Introduction

1.1. Macromolecules and Their Function

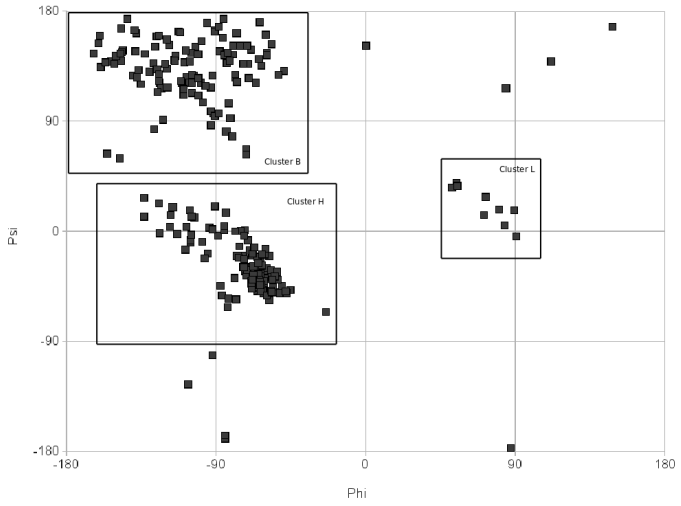
The function of macromolecules is determined by their three-dimensional (3D) structure. This 3D structure allows for a specific binding of small compounds like drugs, metabolites, or other macromolecules such as RNA and proteins. This binding process is crucial for cell signaling and of great interest for understanding the cellular apparatus and the development of new treatments for diseases. Determining the structure of a macromolecule (protein, RNA) and thus the coordinates of the residues in atomic detail was and still is a significant procedure. The first structures, hemoglobin and myoglobin, were determined 1958 by Kendrew *et al.* [1]. Since then progress has been made towards a faster determination of macromolecular structures. However, for many macromolecules it is still impossible to determine the complete structure [2]. The principal repository for the coordinates of macromolecular structures is the wwPDB archive [3]. As of November 2008 the wwPDB stores well over 50,000 structures consisting of roughly 1,500 RNA structures including protein - RNA complexes and 48,000 proteins. In recent years various structural genomics initiatives were started that aimed towards a fast, high-density determination of thousands of macro-molecular structures [4, 5]. These initiatives led to around 1,500 structures with unknown functions. The annotation of macromolecules can be carried out on different levels, however, the manual annotation of those structures is often not feasible despite best efforts [6, 7]. The fastest way to determine the function is to use the sequence of its building blocks (the primary structure) alone, and search this sequence against a database of annotated structures where the function can be subsequently inferred. This approach generally works well when the sequences are highly similar but sometimes fails [8]. A more accurate way to annotate is to use 3D information. Many methods try to identify secondary structure elements and align them with each other. These approaches are often sequence-independent and therefore not subject to failure because of relative sequence similarity [9, 10]. A general fact for both protein and RNA structural alignment is that there often cannot be a single best solution to align two or more structures. The best solution is always the best given a certain man-made optimization criteria, nicely explained by [11].

1.2. Protein Function and Similarity

The importance of structural alignments of protein structures is based on the fact that structural motifs (folds) contained in the structure reveal important biochemical functions [7]. For instance the so-called ”Rossman fold” is a strong indication for the binding of nucleotide derivatives [12]. For performance reasons, many computational algorithms work on the sequence level, while also taking into account the 3D secondary structure as guidance [10, 13]. In many scenarios this approach proves to be fast and accurate enough. However, given the already mentioned fact that a similar sequence does not necessarily mean a structural similarity there are a growing number of approaches that use pure 3D information

to overcome this disadvantage [14, 15]. In this regard the authors want to especially stress the SSM project, which is the first software fast enough to search the whole PDB within minutes with a high accuracy based on an abstraction of the 3D structure [16]. Wikipedia currently lists more than 50 different approaches for protein alignment (http://en.wikipedia.org/wiki/Sequence_alignment_software). A detailed comparison of algorithms and approaches in the field is presented by [17, 18]. The approach presented in this work can be adjusted regarding speed and precision / coverage. A schema frequently used to express the backbone of a protein or RNA is to use torsion angles between a well-defined set of atoms. The torsion angles between consecutive amino acids became famous when Ramachandran et al. published the analysis of the ϕ (phi) and ψ (psi) torsion angles (Definition 2) of protein chains in 1963 [19]. Ramachandran showed that the usage of ϕ and ψ angles allows for a clear separation of secondary structure elements (Figure 1). This in turn allows us to judge whether amino acids belong to a certain class of secondary structure elements like α-helices or β-sheets. This notion was frequently applied in the abstraction and search of similar protein structures and is often used together with techniques such as suffix trees and suffix arrays (among others [20–24]). An interesting approach in the field of protein-protein interaction is proposed by Günther *et al.* where known motifs of interacting domains are used to predict potential interactions of novel proteins [25].

Figure 1. The Ramachandran (ϕ-ψ torsion angle) plot of a Thymidylate Synthase (PDB-ID: 1AXW). The cluster in the upper left corresponds to β-sheets, the cluster in the middle left corresponds to α-helices, and the small cluster in the middle right represents left handed helices. The main clusters (B,H,L) are used to translate a protein structure into a string.



1.3. RNA Function and Similarity

In recent years, RNA gained attention due to the discovery of their heavy involvement in the regulatory apparatus of the cell [26]. Apart from the fact that a relatively small amount of RNA structures are contained in the wwPDB they are nevertheless of growing importance [27, 28]. As this field is relatively young, there are only a few structural RNA alignment methods available [20, 29, 30], but interest in the

structure of RNA is rapidly growing. It has to be noted that there is currently no methodology available that allows for the querying of an RNA motif against all RNA structures in real time, as it is provided by SSM for the world of proteins. A schema to express the backbone of an RNA is the usage of η (eta) and θ (theta) pseudotorsion angles (Definition 1) - representing each nucleotide in a chain by two angles. In a thorough analysis of this pseudotorsion representation, eight main classes of conformations have been identified, and this information could be exploited to highlight important features of the RNA structures [31]. A more detailed approach was taken by Richardson et al. where a set of 46 nucleotide conformations is determined based on the seven torsion angles present in a ribose-to-ribose (suite) unit [32]. This representation implicitly includes the pucker of the ribose, and it is detailed enough to track down the conformations in local motifs such as GNRA tetraloops. To do this, however, it is necessary to have well-resolved RNA structures available. Both of these high level abstractions are limited to the RNA backbone, and their accuracy is not sufficient to reconstruct an RNA structure from the string representation alone. Nevertheless, adding information such as canonical and noncanonical base pairs, as well as base stacking provides sufficient input to assemble RNA tertiary structures from such a combined descriptor alone. Recent progress in the field of RNA structure prediction demonstrates the feasibility of this approach [33], except that the attempt to write the structural descriptor as a string has not been made. The RNA Ontology Consortium is currently standardizing the component descriptors of RNA chains in order to facilitate further work on the subject [34].

1.4. Scope of this Work

The aim of this work is to propose a novel approach for the fast hierarchical search of similarities in thousands of macromolecular structures. The method is based on a fast index structure, derived from the field of classical string alignment[35]. But unlike classical sequence-based search methods, the strings can represent structural features of the 3D structure and are sequence independent. Thus, this approach has the potential to be as fast as sequence-based approaches with the precision of structural alignment methods. The authors want to stress the term ”fast,” as many approaches currently work in a one-against-one mode. The proposed method, materialized in the LaJolla framework, is easily extendable with new chain-to-string translators. The aim of this publication is to present this approach and the software package as a potentially useful tool for many domains. The in-depth validation for individual domains such as protein and RNA similarity, protein-protein interaction and protein-small compound docking is subject to publications in journals of the corresponding communities.

2. Material and Methods

2.1. In a Nutshell

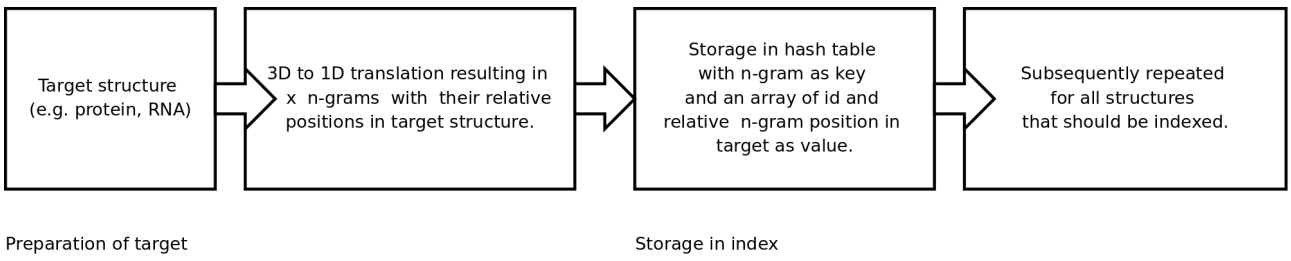
The proposed approach performs a search for local structural motifs in a set of 3D structures of macromolecules. The basic ideas for this approach originate in [36], where we analyzed different possibilities to represent RNA structures as a reduced alphabet and the possibility of storing and querying that alphabet in suffix-based indices. The paper clearly shows that the proposed methodology of suite codes [32] is too narrow for a RNA search. To tackle this drawback we used the notion of n-grams. However, n-grams can be used more easily in hash tables than in the originally implemented suffix-based structures. We

were also interested in how the sophisticated suite methodology compares to a simple η-θ torsion angle discretization. A novel development that enhances the practical application is that LaJolla performs a final 3D superposition to remove statistical artifacts that have no real 3D significance. During the development, it turned out that the approach is not only useful for RNA structures but also for proteins using the respective transformers (for instance ϕ-ψ).

String representation of linear polymers

This approach is based on the simple observation that macromolecular structures share a common property: They are made up of chains formed by molecular building blocks, and possess a linear molecular backbone with repeating units. This property allows for the application of abstractions that are able to translate these macromolecules into a one-dimensional (1D) linear representation (Figure 2). This in turn allows for the use of efficient algorithms deriving from the field of string matching and text mining [35].

Figure 2. An informal diagram showing the steps issued for the initial generation of the hash table used as index structure for the structural alignment.



From torsion angles to a string

A simple translation procedure from a 3D structure to a 1D string is to use the sequence of nucleic acids or amino acids. However, as already stated in the introduction, a high similarity in sequence does not imply that the structures are similar. To overcome this, the default procedure uses torsion angles between defined atoms. In the case of RNA structures, the translator maps the residues to η-θ pseudotorsion angles (Definition 1). In case of proteins ϕ-ψ torsion angles are used (Definition 2).

Definition 1 *Given three consecutive nucleotides N1, N2, N3 of a nucleic acid chain. Let η be the torsion angle defined by atoms ($N1_{C4'}$, $N2_P$) and ($N2_{C4'}$, $N3_P$). Let θ be the torsion angle defined by atoms ($N2_P$, $N2_{C4'}$) and ($N3_P$, $N3_{C4'}$).*

Definition 2 *Given three consecutive amino acids A1, A2, A3 of a polypeptide chain let ϕ be the torsion angle defined by atoms ($A1_C, A2_N$) and ($A2_{CA}$, $A2_C$). Let ψ be the torsion angle defined by atoms ($A2_N, A2_{CA}$) and ($A2_C$, $A3_N$).*

Once a sequence of torsion angles is generated it can be translated into a sequence of characters using any function. In the case of proteins as well as for RNAs the main clusters of the dihedral angle plots are translated into distinct characters (see also Figure 1).

The result of that translation step is a string where single characters represent the torsion angles of the chain residues and therefore the macromolecule as a whole. Traditional string matching algorithms can subsequently be applied, enabling the user to index and to search for macromolecular structures.

An n-gram based index structure for fast searches

A hash table is a data structure that stores key - value pairs. A value can be a character, a string or an arbitrary object. The key is generated by a mathematical function (hashing function) that translates the value into the key. This key in turn allows us to retrieve the value from a hash table in an average run time of O(1) [37]. There are two characteristics of hash tables that have major influence on the run time. First, not all hashing functions necessarily yield unique results, subsequently, collisions have to be resolved by chaining values or by other approaches. Second, to obtain the average run time of O(1) an average load factor has to be kept, and a so-called rehashing has to be issued if the load factor goes below a certain threshold. A good general introduction to the field is given by [38]. To conclude, a hash table allows for a fast determination if certain strings are contained in the index. However, storing the complete sequence of a chain (e.g. discrete η- θ values) as value in the hash table does not make much sense because it would only allow searches for exact matches of whole structures that virtually never occur. To overcome this disadvantage it is useful to store so called n-grams (also: q-grams) of a sequence in the hash table [39]. An n-gram is a string of length n. All n-grams of a string m are all sub-strings of length n of m. For example all 2-grams of the string ALICE are AL, LI, IC and CE. N-grams are widely used as a statistical tool to define the relatedness of two strings. Google's "Did you mean: ..." feature is a classic example of that. But n-grams can also be used as method for fuzzy string alignment. If the string ALICE is searched in the string ALITE using 2-grams, then two 2-grams are found, two missed, and an alignment can be proposed by this approach.

2.2. Generating and Searching the Index

For searching, a hash table is generated from all n-grams of all target structures, in which n-grams generated from the query structure are searched. Generating the n-gram based index is a straightforward process (Figure 2). All target structures (chains) have to be translated subsequently to strings using a structure-to-string translator. The n-grams of each target structure are stored in the hash table. It has to be noted that the positions of the n-grams of query and target are stored as well, making it feasible to perform a 3D alignment for scoring and refinement. Searching a structure (query) in the index involves the transformation of the query chain into a string and the computation of each n-gram (Figure 4). The search results in a certain amount of target structures that have n-grams in common with the query. As these results may be statistical artifacts, a second hierarchical refinement step is applied. In this refinement step, the corresponding n-grams are subsequently aligned and thus anchors of query and target are determined. With that allocation, a superposition of query and target is performed [40]. The scoring is carried out by calculating the RMSD (Definition 3) and a qualitative score, TM-Score, as defined in [41] (Definition 4). The RMSD alone is not suitable as it does not allow conclusions to be drawn about the number of residues that have been aligned successfully.

Definition 3

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2} \tag{1}$$

where d_i is the Euclidian distance between N pairs of equivalent residues. The RMSD is calculated in Ångström.

Definition 4

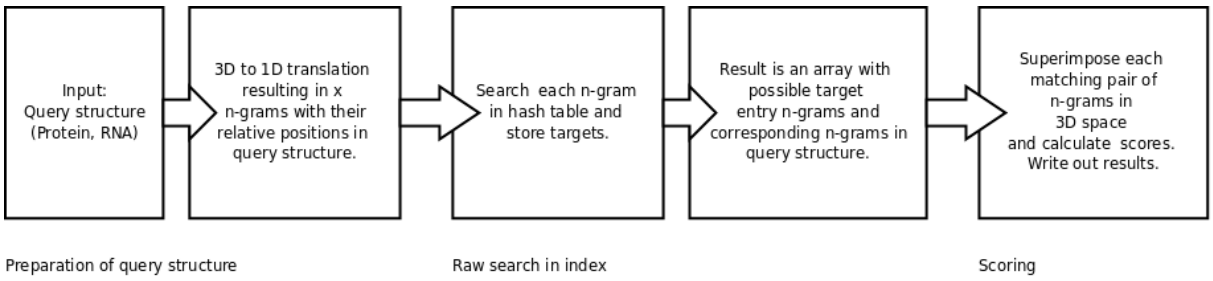
$$\mathrm{TM-Score} = \frac{1}{L_{Target}} \sum_{i=1}^{L_{aligned}} \frac{1}{1 + \left(\frac{d_i}{1.24\sqrt[3]{L_{Target}-15}-1.8}\right)^2} \tag{2}$$

where L_{Target} and $L_{aligned}$ are the lengths of the target and aligned structure respectively. d_i is the Euclidian distance between the ith pair of residues.

A conceptional advantage of the presented approach is that the strings (n-grams) that are being indexed and in turn searched using the hash table can be generated by an arbitrary approach. From the perspective of software engineering it is easily possible to exchange the discussed approach of protein ϕ-ψ torsion with the Protein Blocks Method [42] mentioned in the introduction. For RNA structures it would be easily possible to replace the η-θ torsion angles approach with the notion of suite codes proposed by Richardson *et al.* [32], or any other representation.

The principal parameters that have an impact on performance and accuracy are the size (n) of the indexed n-grams and complexity of the string a structure to string translator produces. In an extreme case a structure to string translator would produce always the same letter for each angle combination meaning each n-gram of the query will be compared to the each n-gram of the target. A clever translator reduces this by only comparing beta-sheets and helices or even combinations using a longer n-gram size reducing the search time dramatically.

Figure 3. An informal diagram showing the steps performed when a query structure is searched against a set of target structures in an index.



2.3. Datasets Used

The datasets can be downloaded from the project homepage at http://lajolla.sf.net.

tRNA dataset

For the analysis of the RNA alignment capabilities of LaJolla, all molecular structures containing a tRNA were retrieved from the NDB database. The dataset was filtered manually, to identify the polymer chains, to identify the functional state of the molecules, and exclude structural fragments. The resulting dataset contains 101 nucleic acid chains, all of which have been resolved by x-ray crystallography.

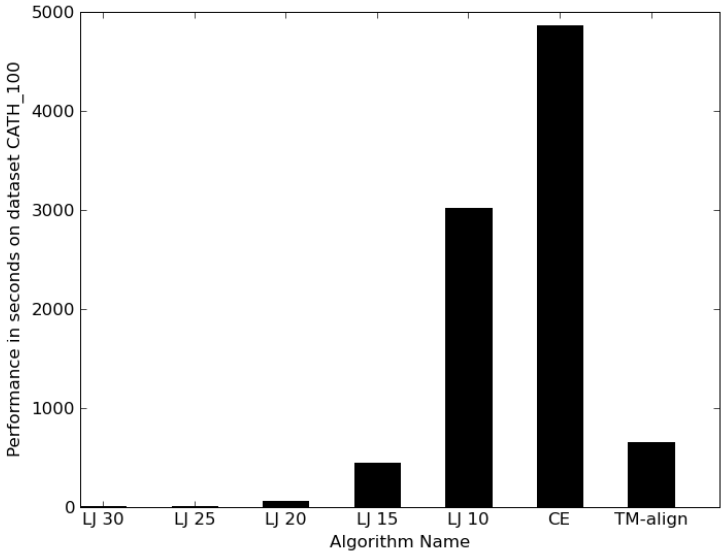
Protein benchmarking

We used two datasets for benchmarking LaJolla in the protein domain. The first Dataset termed CATH_1258 is derived from the CATH [43]. We are using the S35 subset of CATH version 3.2.0 (CathDomainPdb.S35). From this subset we picked the first entry of each structure at "H" level. We subsequently removed entries that are singletons regarding their parent topology level resulting in 1,258 entries. Thus, each of these 1,258 structures is classified by CATH and has at least one entry that is classified in the same class - architecture - topology combination. For the evaluation of the performance we randomly picked 100 structures from the S35 subset of CATH termed CATH_100.

3. Results

The following results were performed with the default settings of LaJolla version 2.0. Results below a TM-Score of 0.2 are neglected. The standard translators were used, for proteins *BetterOptimizedPhiPsiTranslator*, for RNA structures *OptimizedStructureToEtaThetaCharacterTransformer*. For comparison to the state of the art we used CE (version 2004/10/07) [10] and TM-align (64bit version 2005/06/01) [13].

Figure 4. An informal diagram showing the steps performed when a query structure is searched against a set of target structures in an index.



3.1. Performance

As the CATH_1258 dataset was executed on a distributed environment it is not possible to take these runtimes. To this end, the CATH_100 dataset was used and executed in an all against all manner for LaJolla (n-gram sizes 10, 15, 20, 25, 30), CE and TM-align (Figure 4). The tests were executed on standard hardware with an Opteron 2.2 GHz with only one CPU enabled. The histogram points out that LaJolla is fast when using larger n-gram sizes. CE is the slowest method.

3.2. RNA Retrieval

A multiple structural alignment of 101 tRNA chains was performed using the η-θ angle representation implemented in LaJolla. The tRNA molecule was chosen for this task, as it possesses a highly conserved tertiary structure that is straightforward to recognize and to validate. Despite that, it contains many local structural variations, and changes its conformation depending on its functional state (see [44] for a review). Finally, a high number of structures of different quality are available for this family of RNA. The all-against-all search in LaJolla resulted in $101^2 = 10,201$ queries that were performed with an n-gram size of 10. The run resulted in 10,195 local alignments returned by the program. To validate the results, it was checked how well the query and target structures are superimposed by the method. By manual inspection it was determined that finding at least 30 residues close to each other, or a TM-score higher than 0.25, were sufficient criteria to tell apart correct global superpositions and mere local similarities. Using these criteria, 9,237 (90.5%) superpositions were done successfully.

The full list of examined RNA chains and average RMSD, TM-score, and number of aligned residues are given in Tables 1 and 2. Inspecting the results in detail, it was found that for the RNA chains 2nre/F and 1j2b/C+D more than 60% of the superpositions failed. In both cases, the RNAs are forced by a base modifying enzyme into an unusual conformation (pseudouridine synthase and archaeosine transglycosylase, respectively). In the case of 1j2b, an entire arm of the tRNA changes its conformation (called lambda-form tRNA). Other functional states of the matched tRNA molecules shows little influence on the number of hits. By far the most abundant state available is tRNA bound to aaRS proteins (56 chains), and it has on average 91.8 correct hits found by LaJolla. The next most frequent group are ribosomes (26 chains), with 96.3 correct hits in average. In total, there are on average 91.5 correct hits per chain. The number of alignments found may result from similarities of the functional states, but we were not able to confirm this as significant - for this, one would expect e.g. tRNA in complex with aaRS to prefer each other in the hit list, and ribosome complexes among each other etc. We do not, however, observe this. A bad resolution seems to rather improve the alignability of a structure, as observed on the ribosomes. A simple explanation for this is that the tRNA in many of the ribosomal structures has been constructed by molecular recognition techniques using a standard template - and intricate local variations not detectable in the structures. As a result, it can be stated that, using the default η-θ translator, it is possible to align badly-resolved structures correctly - a feature not attainable by the suite code translator. On the positive side, inspection of the local alignments showed that they are not altogether local. LaJolla finds a series of matching n-grams throughout a pair of structures. Thus, the structural alignments are not based on a local similar substructure common to both molecules, but rather a consensus of many small similarities

that add together to the final alignment. Only in the incorrect hits was the alignment confined to some part of the structure.

For analyzing the sensitivity of the tRNA structural alignment, we compared 60 RNA structures annotated in the SCOR database, including 13 tRNAs (taken from [30]). It was calculated, how many times the highest scoring structure retrieved by LaJolla has the same class in the 'functional annotation' category. For tRNA, this was the case for 100% of the entries regardless of n-gram size. This shows that tRNA structures being that similar to each other that even a moderately accurate superposition it sufficient to distinguish them from other types of RNA. When considering the accuracy of other functional classes, the retrieval gets much less accurate, with only 53% correctly assigned functional categories (when considering the best of the top five TM-scores, this number rises to 69%). One of the reasons for the observed wrong assignments is that part of the 47 non-tRNA structures express considerable structural variety despite their small size. This sensitivity can be improved by applying a TM-score cutoff, but this may lead to misleading figures because then the tRNA structures will be heavily overrepresented in the data. This points to limits of the SCOR dataset, and suggests that a manual functional annotation of those parts of the PDB not covered by SCOR would be helpful.

3.3. Protein retrieval

To evaluate the capabilities of LaJolla in the field of protein retrieval we used the CATH protein classification as standard of truth and compared the results to two other popular algorithms in the field: CE [10] and TM-align [13]. CATH allows us to validate if the results produced by a method are "true" in terms of a similar classification. To this end, we used the topology level of CATH. Our reduced dataset CATH_1258 ensures that there is at least one other protein on the same topology level. The graphs in Figure 5 show how well the classification works in regard to the coverage at a certain scoring cutoff. To assess correct hits, we counted if the result with the best score was true (TOP 1) and also if a correct result was among the ten best hits (TOP 10). The results show that CE, despite its age, still is a very good method with good overall results. It ranks best when it comes to TOP 1 hits and second when it comes to TOP 10 hits. If one takes into account when the coverage line crosses the percentage of correct TOP1 and TOP 10 hits CE also ranks first. TM-align is faster than CE, and has the best characteristics regarding the TOP 10 hits in the field when considering a TM-Score between 0.0 and 0.5. LaJolla's coverage and sensitivity can be adjusted using the n-gram size. The TOP 1 hits with n-gram length 10 are equally good as the results produced by TM-align. The results of LaJolla show that a certain amount of chain gets lost when using longer n-gram sizes as they cannot be indexed. As LaJolla was used with standard parameters results below a TM-Score of 0.2 were neglected what also contributes to this. However, as the performance graph shows (4), this is a tradeoff between speed and coverage / precision. The performance is higher compared to CE and TM-align in all n-gram sizes except 10 where TM-align is faster.

Figure 5. Evaluation of the coverage and precision of LaJolla (n-gram size 10, 20, 25, 30), CE and TM-align in a classification scenario. The red line indicates the percentage of the coverage of distinct topologies at a certain score cutoff. The black line represents the percentage of correct hits with the best score (TOP 1), the dashed black line represents the percentage of correct assignments with a true result being among the ten best hits (TOP 10).

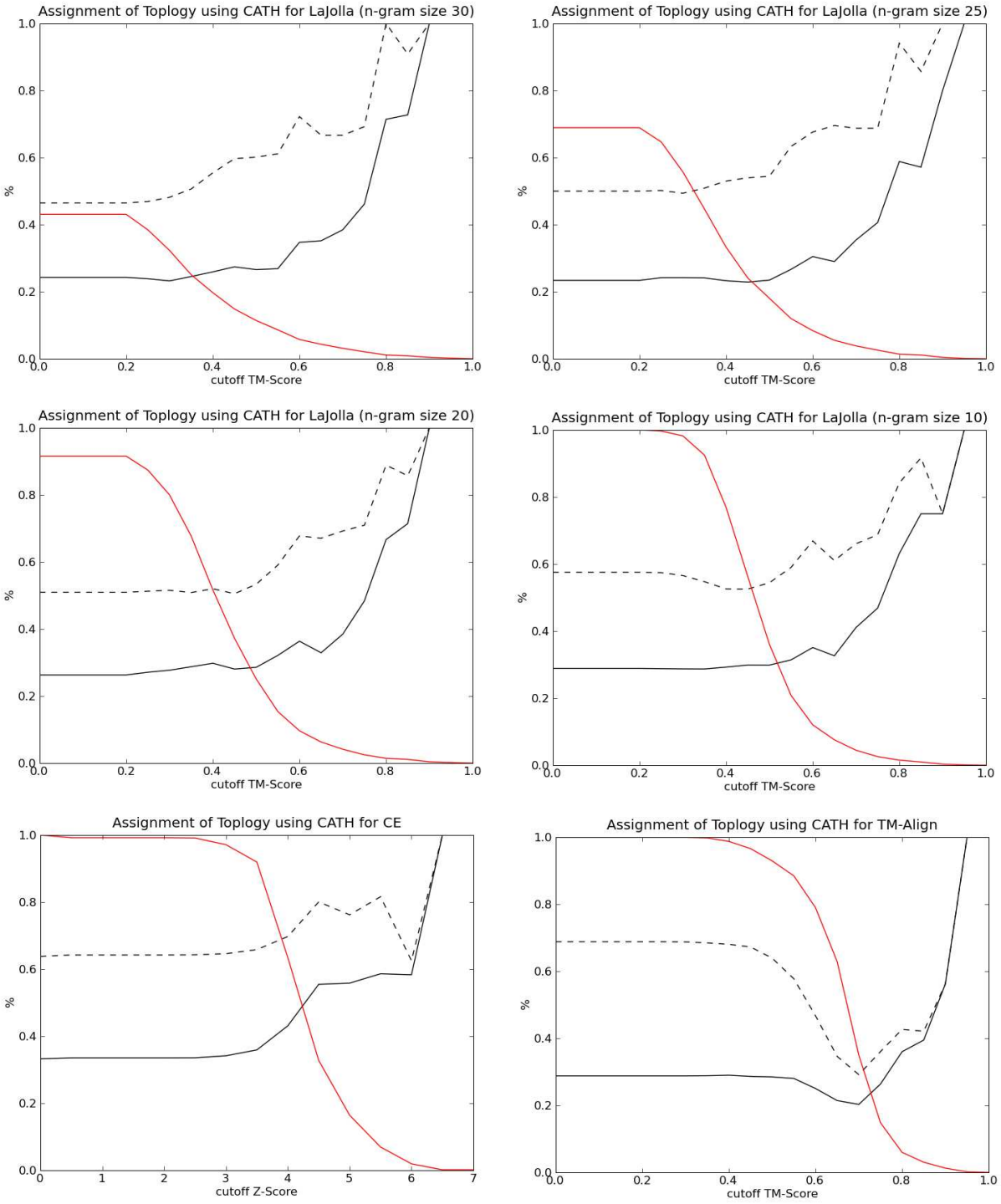


Table 1. tRNA search result table (part 1).

PDB-ID	chain	resolution	tRNA type	complex with	N hits	RMSD	TM-Score	percentage aligned residues
1b23	R	2,60	tRNA_Cys	Ef-Tu	87	1.75	0.38	36.24
1c0a	B	2,40	tRNA_Asp	AspRS	98	1.71	0.48	44.85
1efw	C	3,00	tRNA_Asp	AspRS	92	1.76	0.48	44.95
1efw	D	3,00	tRNA_Asp	AspRS	95	1.73	0.48	44.29
1ehz	A	1,93	tRNA_Phe	uncomplexed	98	1.70	0.52	49.18
1eiy	C	3,30	tRNA_Phe	PheRS	65	1.89	0.33	33.23
1euq	B	3,10	tRNA_Gln	GlnRS	98	1.71	0.52	46.91
1euy	B	2,60	tRNA_Gln	GlnRS	98	1.67	0.52	47.00
1exd	B	2,70	tRNA_Gln	GlnRS	99	1.75	0.51	47.07
1f7u	B	2,20	tRNA_Arg	ArgRS	98	1.75	0.43	40.52
1f7v	B	2,90	tRNA_Arg	ArgRS	98	1.74	0.44	40.66
1ffy	T	2,20	tRNA_Ile	IleRS	96	1.69	0.48	44.37
1g59	B	2,40	tRNA_Glu	GluRS	89	1.62	0.49	44.88
1g59	D	2,40	tRNA_Glu	GluRS	88	1.65	0.49	44.54
1gts	B	2,80	tRNA_Gln	GlnRS	95	1.70	0.48	44.53
1h3e	B	2,90	tRNA_Tyr	TyrRS	97	1.77	0.45	42.83
1h4s	T	2,85	tRNA_Pro	ProRS	91	1.68	0.45	38.62
1il2	C	2,60	tRNA_Asp	AspRS	90	1.84	0.45	44.02
1il2	D	2,60	tRNA_Asp	AspRS	96	1.72	0.47	42.05
1j1u	B	1,95	tRNA_Tyr	TyrRS	99	1.63	0.50	45.30
1j2b	C	3,30	tRNA_Val	archaeosine transglycosylase	38	1.79	0.34	32.84
1j2b	D	3,30	tRNA_Val	archaeosine transglycosylase	31	1.80	0.35	32.43
1n77	C	2,40	tRNA_Glu	GluRS	94	1.62	0.49	44.59
1n77	D	2,40	tRNA_Glu	GluRS	91	1.68	0.50	46.08
1n78	C	2,10	tRNA_Glu	GluRS	93	1.62	0.50	45.50
1n78	D	2,10	tRNA_Glu	GluRS	91	1.69	0.50	46.42
1ob2	B	3,35	tRNA_Phe	Ef-Tu	97	1.85	0.43	42.31
1pns	V	8,70	tRNA_Phe	70S ribosome	98	1.71	0.53	49.57
1pns	W	8,70	tRNA_Phe	70S ribosome	99	1.70	0.50	46.53
1qf6	B	2,90	tRNA_Thr	ThrRS	96	1.71	0.45	41.68
1qrs	B	2,60	tRNA_Gln	GlnRS	94	1.69	0.49	45.49
1qrt	B	2,70	tRNA_Gln	GlnRS	94	1.70	0.48	44.62
1qru	B	3,00	tRNA_Gln	GlnRS	94	1.69	0.49	44.97
1qtq	B	2,25	tRNA_Gln	GlnRS	98	1.68	0.49	44.82
1qu2	T	2,20	tRNA_Ile	IleRS	96	1.69	0.48	44.37
1qu3	T	2,90	tRNA_Ile	IleRS	98	1.68	0.49	44.93
1wz2	C	3,21	tRNA_Leu	LeuRS	97	1.75	0.43	40.84
1wz2	D	3,21	tRNA_Leu	LeuRS	97	1.74	0.46	43.21
1yl4	B	5,50	tRNA_Phe	70S ribosome	98	1.83	0.50	48.26
1yl4	C	5,50	tRNA_Phe	70S ribosome	99	1.75	0.50	47.07
1zjw	B	2,50	tRNA_Glu	GluRS	98	1.68	0.50	45.61
2ake	B	3,10	tRNA_Trp	TrpRS	96	1.67	0.44	40.14
2azx	C	2,80	tRNA_Trp	TrpRS	100	1.72	0.50	45.71
2azx	D	2,80	tRNA_Trp	TrpRS	100	1.73	0.48	44.01
2b64	V	5,90	tRNA_Phe	70S ribosome	98	1.76	0.47	45.16
2b64	W	5,90	tRNA_Phe	70S ribosome	98	1.82	0.52	49.75
2b9m	V	6,76	tRNA_Phe	70S ribosome	98	1.77	0.47	44.88
2b9m	W	6,76	tRNA_Phe	70S ribosome	99	1.83	0.48	46.98
2b9o	V	6,46	tRNA_Phe	70S ribosome	100	1.78	0.46	44.43
2b9o	W	6,46	tRNA_Phe	70S ribosome	98	1.79	0.51	49.07
2bte	B	2,90	tRNA_Leu	LeuRS	86	1.88	0.42	40.93
2bte	E	2,90	tRNA_Leu	LeuRS	81	1.84	0.41	40.17
2byt	B	3,30	tRNA_Leu	LeuRS	72	1.85	0.41	40.15
2byt	E	3,30	tRNA_Leu	LeuRS	71	1.85	0.42	40.36
2csx	C	2,70	tRNA_Met	MetRS	95	1.68	0.47	44.03
2csx	D	2,70	tRNA_Met	MetRS	95	1.66	0.47	43.39
2ct8	C	2,70	tRNA_Met	MetRS	99	1.69	0.47	43.53
2ct8	D	2,70	tRNA_Met	MetRS	97	1.70	0.43	40.05
2cv0	C	2,40	tRNA_Glu	GluRS	93	1.62	0.49	44.53
2cv1	C	2,41	tRNA_Glu	GluRS	93	1.64	0.50	45.99
2cv1	D	2,41	tRNA_Glu	GluRS	91	1.70	0.50	46.78
2cv2	C	2,69	tRNA_Glu	GluRS	92	1.65	0.51	46.75
2cv2	D	2,69	tRNA_Glu	GluRS	91	1.69	0.50	46.31
2d6f	E	3,15	tRNA_Gln	GluRS	97	1.80	0.43	40.94
2d6f	F	3,15	tRNA_Gln	GluRS	98	1.87	0.41	40.11
2der	C	3,10	tRNA_Glu	mnma thiolase	98	1.74	0.48	44.98
2der	D	3,10	tRNA_Glu	mnma thiolase	96	1.70	0.50	44.92
2det	C	3,40	tRNA_Glu	mnm5s2U-methyltransferase	94	1.72	0.45	40.28
2deu	C	3,40	tRNA_Glu	mnm5s2U-methyltransferase	90	1.73	0.43	40.93
2deu	D	3,40	tRNA_Glu	mnm5s2U-methyltransferase	89	1.73	0.44	41.02
2dr2	B	3,00	tRNA_Trp	TrpRS	100	1.68	0.43	39.79
2du3	D	2,60	tRNA_Cys	o-phosphoserylRS	95	1.76	0.45	41.51
2du4	C	2,80	tRNA_Cys	o-phosphoserylRS	95	1.78	0.46	42.32
2du5	D	3,20	tRNA_opal	o-phosphoserylRS	93	1.88	0.41	39.22
2du6	D	3,30	tRNA_Amber	o-phosphoserylRS	96	1.89	0.40	38.27
2dxi	C	2,20	tRNA_Glu	GluRS	92	1.62	0.49	44.98
2dxi	D	2,20	tRNA_Glu	GluRS	88	1.63	0.49	44.78
2fk6	R	2,90	tRNA_Thr	RNase Z	85	1.57	0.52	36.29

Table 2. tRNA search result table (part 2).

PDB-ID	chain	resolution	tRNA type	complex with	N hits	RMSD	TM-Score	percentage aligned residues
2hgi	C	5,00	tRNA_fMet	70S ribosome	99	1.69	0.52	48.56
2hgi	D	5,00	tRNA_Phe	70S ribosome	86	1.90	0.42	41.56
2hgp	B	5,50	tRNA_Phe	70S ribosome	90	1.91	0.44	43.47
2hgp	C	5,50	tRNA_Phe	70S ribosome	98	1.77	0.49	46.43
2hgp	D	5,50	tRNA_Phe	70S ribosome	90	1.84	0.42	41.07
2hgr	C	4,51	tRNA_fMet	70S ribosome	100	1.68	0.51	47.38
2hgr	D	4,51	tRNA_Phe	70S ribosome	93	1.89	0.43	42.78
2iy5	T	3,10	tRNA_Phe	PheRS	51	1.95	0.33	33.58
2j00	W	2,80	tRNA_Phe	70S ribosome	97	1.76	0.44	42.02
2j02	V	2,80	tRNA_fMet	70S ribosome	98	1.69	0.49	46.10
2j02	W	2,80	tRNA_Phe	70S ribosome	97	1.80	0.46	44.09
2nre	F	4,00	tRNA_Leu	pseudouridine synthase	32	1.56	0.46	33.68
2ow8	0	3,71	tRNA_Phe	70S ribosome	93	1.89	0.42	41.68
2ow8	z	3,71	tRNA_Phe	70S ribosome	90	1.82	0.45	43.30
2qnh	2	3,83	tRNA_Phe	70S ribosome	93	1.84	0.43	41.65
2qnh	z	3,83	tRNA_fMet	70S ribosome	100	1.74	0.51	48.56
2tra	A	3,00	tRNA_Asp	uncomplexed	98	1.72	0.44	40.72
2v0g	B	3,50	tRNA_Leu	LeuRS	69	1.86	0.42	40.74
2v0g	F	3,50	tRNA_Leu	LeuRS	69	1.85	0.42	40.22
2v46	W	3,80	tRNA_fMet	70S ribosome	98	1.80	0.46	44.24
2v48	W	3,80	tRNA_fMet	70S ribosome	96	1.86	0.46	44.87
3tra	A	3,00	tRNA_Asp	uncomplexed	93	1.75	0.45	41.92
4tna	A	2,50	tRNA_Phe	uncomplexed	100	1.70	0.52	49.00

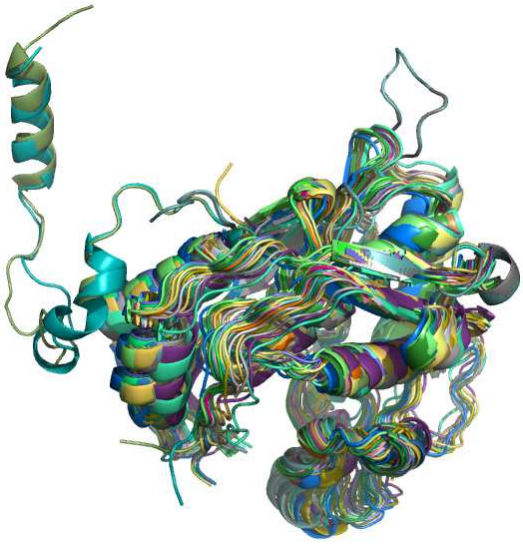
4. Discussion

4.1. General aspects

The aims of this approach as defined in the introduction were the proposition of a generalized methodology that can be extended and customized by the user for different macromolecules and applications. In the results section we showed that the performance and precision / coverage of the approach is comparable to common methods available freely today. The trade-off between performance and precision / coverage can be adjusted using the n-gram length. The described chain to string translators are independent from an initial precomputation of the secondary structure elements. With the dataset CATH_1258 derived from the 35% filtered CATH it becomes clear, that the approach works well when the sequence of the proteins is not entirely similar. Moreover, because this approach is implemented as open source in the framework LaJolla, it can be easily extended with novel translators that abstract the macromolecular structure in different ways such as suite codes or protein building blocks. It has to be pointed out that the results presented for both proteins and RNA were achieved by the backbone information alone. It is safe to assume that on both sides the accuracy of the approach could be improved by including sequence-specific information. In the case of RNA, this could be for instance the isostericity matrices of Leontis and Westhof [45]. The principal performance bottleneck is the refinement step where the biomolecules have to be read from the hard disk and superimposed in 3D. Almost 80% of the time currently used for search are input / output operations. It is possible to tackle this problem from many sides. The implementation of a caching infrastructure that stores frequently used structures in memory so that subsequent hard disk reads are redirected to memory would be the first logical step. Another possibility is to store specially prepared files that only contain atoms used for 3D refinement, which would reduce the file size that has to be read. Defining a threshold of how many matching n-grams between query and target at least have to be found to carry out the expensive 3D alignment has the potential to eliminate impossible

alignments beforehand. Although the method was not planned to be used as tool for multiple alignment it can be used for this purpose, by the simple fact that the query structure is never translated / rotated. Subsequently, all target structures are superimposed in a multiple alignment fashion (Figure 6).

Figure 6. A multiple alignment produced by LaJolla using chain B of Thymidylate Synthase with PDB ID 2TSC against all other Thymidylate Synthases.



4.2. RNA specific aspects

The sensitivity of tRNA structural alignment is satisfactory (90.5%). In most cases, where the alignment fails, this is due to drastic structural differences, for instance in the case of lambda-tRNA, where an entire arm of the tertiary structure is displaced by an enzyme. We think that a careful refinement of the parameters (n-gram size, TM-score threshold for alignment) could gain a few percent and superimpose a few additional examples successfully. More worthwhile to try is to run the algorithm on a vast set of RNA structures elucidating how well smaller and bigger types of RNA can be recovered. Such a study should answer how accurate the function of RNA can be recognized in general. A prerequisite for this is a careful and complete functional characterization of RNA structures that does not exist at present. Further, it could be examined whether choosing a different string representation (e.g. Richardson's suite codes) could accelerate the alignment process. But in order to not losing too much sensitivity, the n-gram search would need to account for partial similarity instead of using dissimilarity of two characters as an absolute exclusion criterion. For such and related studies, the tRNA dataset presented here provides a reasonable benchmark that could be used to compare structural search and alignment methods for RNA.

4.3. Protein specific aspects

Using the CATH as standard of truth is generally disputed. TM-align and other algorithms [13, 46] that are originating from the field of protein structure prediction try to score a method based on the coverage of the sequence. This omits the problem that man made classification schemes such as CATH

may contain wrong classifications. However, as LaJolla works completely sequence independent it is not easy to translate the meaning of the results. We therefore used the CATH classification approach as used by other contributions [18]. As we are comparing the results of LaJolla to CE and TM-align this gives a good general view of the capabilities, strengths and weaknesses of the algorithms as possibly wrong classifications are a problem for all algorithms. This methodology also allows the user to judge how to treat results with a certain score. Another general problem is that TM-align and CE do not write out protein structure positions on hard-disk by default. As LaJolla by default always writes superpositions to the hard-disk this is a clear disadvantage for LaJolla and turning off that feature would increase the performance. Still, LaJolla ranks almost always best in terms of performance even with this disadvantage. This suggests that LaJolla is especially useful when it comes to high throughput experiments, where thousands of proteins should be classified and a certain loss of coverage is regrettable.

5. Conclusions

In this work we presented a generalized approach for the fast search and structural alignment of arbitrary macromolecules. The notion of using an index and performing one-against-all searches is a novelty in the world of RNA. This paper showed that the approach yields structural alignments that agree with biological reality using simple ϕ-ψ / η-θ translators. The described approach has an adjustable coverage and precision based on the desired speed using the n-gram size as parameter. This method will be an important aid in the high throughput functional annotation of proteins and RNA, and will make it feasible to search and test new hypotheses about protein and RNA function in a fast manner. The method has obvious applications to the field of knowledge-based docking of small compounds or even proteins. The implementation of this approach, LaJolla, is easy to extend using custom translators (eg . pure amino acid or nucleic acid sequence-based translators). The authors gladly welcome any recommendations and critiques from the community. LaJolla (including platform-independent binary packages, general development resources and mailing lists) is freely available as open source from: http://lajolla.sf.net.

Acknowledgements

The authors want to thank Oliver Buchtala, Stefan Günther, Aysam Guerler, Patrick May and Thomas Röblitz for absolutely inspiring discussions and help, Marcus Schroeder for letting LaJolla make the first steps on the planet RNA and Rebecca F. Miller for excellent proof reading. The LaJolla project team likes to thank Sourceforge (http://sf.net) for project hosting as well as BioJava (http://biojava.org) [47] and Sun (http://java.net) for providing excellent and open source libraries.

References and Notes

1. Kendrew, J.C.; Bodo, G.; Dintzis, H.M.; Parrish, R.G.; Wyckoff, H.; Phillips, D.C. A three- dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **1958**, *181*, 662–666.
2. Scheerer, P.; Park, J.H.; Hildebrand, P.W.; Kim, Y.J.; Krausz, N.; Choe, H.W.; Hofmann, K.P.; Ernst, O.P. Crystal structure of opsin in its G-protein-interacting conformation. *Nature* **2008**, *455*, 497–502.
3. Berman, H.; Henrick, K.; Nakamura, H.; Markley, J.L. The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucl. Acid. Res.* **2007**, *35*, D301–D303.

4. Service, R.F. Structural biology. protein structure initiative: phase 3 or phase out. *Science* **2008**, *319*, 1610–1613.
5. Levitt, M. Growth of novel protein structural data. *Proc. Nat. Acad. Sci.* **2007**, *104*, 3183–3188.
6. Rother, K.; Michalsky, E.; Leser, U. How well are protein structures annotated in secondary databases? *Proteins* **2005**, *60*, 571–576.
7. Andreeva, A.; Howorth, D.; Chandonia, J.M.; Brenner, S.E.; Hubbard, T.J.; Chothia, C.; Murzin, A.G. Data growth and its impact on the SCOP database: new developments. *Nucl.Acid.Res.* **2008**, *36*, 419–425.
8. He, Y.; Chen, Y.; Alexander, P.; Bryan, P.N.; Orban, J. NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc.Nat.Acad.Sci.* **2008**, *105*, 14412–14417.
9. Cheek, S.; Qi, Y.; Krishna, S.S.; Kinch, L.N.; Grishin, N.V. SCOPmap: automated assignment of protein structures to evolutionary superfamilies. *BMC Bioinformatics* **2004**, *5*, 197.
10. Shindyalov, I.N.; Bourne, P.E. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Engeering* **1998**, *11*, 739–747.
11. Sippl, M.J.; Wiederstein, M. A note on difficult structure alignment problems. *Bioinformatics* **2008**, *24*, 426-427.
12. Rao, S.T.; Rossmann, M.G. Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **1973**, *76*, 241-256.
13. Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucl.Acid.Res.* **2005**, *33*, 2302-2309.
14. Guerler, A.; Knapp, E.W. Novel protein folds and their nonsequential structural analogs. *Protein Sci.* **2008**, *17*, 1374-1382.
15. Ilyin, V.A.; Abyzov, A.; Leslin, C.M. Structural alignment of proteins by a novel topofit method, as a superimposition of common volumes at a topomax point. *Protein Sci.* **2004**, *13*, 1865-1874.
16. Krissinel, E.; Henrick, K. Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr-D-Biol Cryst.* **2004**, *60*, 2256-2268.
17. Kolodny, R.; Koehl, P.; Levitt, M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.* **2005**, *346*, 1173–1188.
18. Novotny, M.; Madsen, D.; Kleywegt, G.J. Evaluation of protein fold comparison servers. *Proteins* **2004**, *54*, 260–270.
19. Ramachandran, G.N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95-99.
20. Guyon, F.; Camproux, A.C.; Hochez, J.; Tuffery, P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. *Nucl.Acid.Sci.* **2004**, *32*, W545-548.
21. Täubig, H.; Buchner, A.; Griebsch, J. PAST: Fast structure-based searching in the PDB. *Nucl.Acid.Sci.* **2006**, *34*, W20-W23.
22. Friedberg, I.; Harder, T.; Kolodny, R.; Sitbon, E.; Li, Z.; Godzik, A. Using an alignment of fragment strings for comparing protein structures. *Bioinformatics* **2007**, *23*, e219-e224.
23. Lo, W.C.; Huang, P.J.; Chang, C.H.; Lyu, P.C. Protein structural similarity search by Ramachandran codes. *BMC Bioinformatics* **2007**, *8*, 307.
24. Gao, F.; Zaki, M.J. PSIST: A scalable approach to indexing protein structures using suffix trees. J.

Parallel Distributed Computation **2008**, *68*, 54-63.

25. Günther, S.; May, P.; Hoppe, A.; Frömmel, C.; Preissner, R. Docking without docking: ISEARCH-prediction of interactions using known interfaces. *Proteins* **2007**, *69*, 839-844.
26. Laederach, A. Informatics challenges in structured RNA. *Brief Bioinformatics* **2007**, *8*, 294-303.
27. Tamura, M.; Hendrix, D.K.; Klosterman, P.S.; Schimmelman, N.R.; Brenner, S.E.; Holbrook, S.R. SCOR: Structural Classification of RNA, version 2.0. *Nucl. Acid. Res.* **2004**, *32*, D182-D184.
28. Abraham, M.; Dror, O.; Nussinov, R.; Wolfson, H.J.J. Analysis and classification of RNA tertiary structures. *RNA* **2008**, *14*, 2274-2289.
29. Chang, Y.F.F.; Huang, Y.L.L.; Chin. SARSA: a web tool for structural alignment of RNA using a structural alphabet. *Nucl. Acid. Res.* **2008**, *36*, 19-24.
30. Capriotti, E.; Marti-Renom, M.A. RNA structure alignment by a unit-vector approach. *Bioinformatics* **2008**, *24*, 112–118.
31. Wadley, L.M.; Keating, K.S.; Duarte, C.M.; Pyle, A.M. Evaluating and Learning from RNA Pseudotorsional Space: Quantitative Validation of a Reduced Representation for RNA Structure. *J. Mol. Biol.* **2007**, *372*, 942–957.
32. Richardson, J.S.; Schneider, B.; Murray, L.W.; Kapral, G.J.; Immormino, R.M.; Headd, J.J.; Richardson, D.C.; Ham, D.; Hershkovits, E.; Williams, L.D.; Keating, K.S.; Pyle, A.M.; Micallef, D.; Westbrook, J.; Berman, H.M. RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **2008**, *14*, 465-481.
33. Parisien, M.; Major, F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **2008**, *452*, 51–55.
34. Leontis, N.B.; Altman, R.B.; Berman, H.M.; Brenner, S.E.; Brown, J.W.; Engelke, D.R.; Harvey, S.C.; Holbrook, S.R.; Jossinet, F.; Lewis, S.E.; Major, F.; Mathews, D.H.; Richardson, J.S.; Williamson, J.R.; Westhof, E. The RNA Ontology Consortium: an open invitation to the RNA community. *RNA* **2006**, *12*, 533–541.
35. Gusfield, D. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, 1997.
36. Bauer, R.A.; Rother, K.; Bujncki, J.; Preissner, R. Suffix techniques as a rapid method for RNA substructure search. *Genome Informatics* **2008**, *20*, 183–198.
37. Dietzfelbinger, M.; Karlin, A.R.; Mehlhorn, K.; Meyer auf der Heide, F.; Rohnert, H.; Tarjan, R.E. Dynamic perfect hashing: Upper and lower bounds. In *IEEE Symposium on Foundations of Computer Science*, 1988, 524–531.
38. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. Introduction to Algorithms. McGraw-Hill Science / Engineering / Math, 2nd Edition, 2003.
39. Burkhardt, S.; Crauser, A.; Ferragina, P.; Lenhof, H.P.; Rivals, E.; Vingron, M. q-gram based database searching using a suffix array (QUASAR). In *RECOMB '99: Proceedings of the third annual international conference on Computational molecular biology* 1999, 77-83.
40. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **1976**, *32*, 922-923.
41. Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **2007**, *57*, 702-710.

42. Tyagi, M.; de Brevern, A.G.; Srinivasan, N.; Offmann, B. Protein structure mining using a structural alphabet. *Proteins* **2008**, *71*, 920-937.
43. Cuff, A.L.; Sillitoe, I.; Lewis, T.; Redfern, O.C.; Garratt, R.; Thornton, J.; Orengo, C.A. The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucl. Acid. Res.* **2009**, *37*, D310-314.
44. Giegé, R. Toward a more complete view of tRNA biology. *Nat. Struct. Mol. Biol.* **2008**, *15*, 1007-1014.
45. Stombaugh, J.; Zirbel, C.L.; Westhof, E.; Leontis, N.B. Frequency and isostericity of RNA base pairs. *Nucl. Acid. Res.* **2009** *in press*.
46. Pandit, S.B.; Skolnick, J. Fr-TM-align: A new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics* **2008**, *9*, 531.
47. Holland, R.C.; Down, T.; Pocock, M.; Prlic, A.; Huen, D.; James, K.; Foisy, S.; Dräger, A.; Yates, A.; Heuer, M.; Schreiber, M.J. BioJava: an open-source framework for bioinformatics. *Bioinformatics* **2008**, *24*, 2096-2097.