Symbiosis

Research Article | Journal of Computer Science Applications and Information Technology | Open Access

Received: March 08, 2017
Accepted: March 20, 2017
Published: April 27, 2017

Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB

Eugene Krissinel[1]* and Ville Uski[2]

[1]CCP4 Core Group Leader, Scientific Computing Department, Didcot, OX11 0FA, United Kingdom
[2]CCP4 Computational Scientist, Scientific Computing Department, Didcot, OX11 0FA, United Kingdom

***Corresponding author:** Eugene Krissinel, CCP4 Core Group Leader, Scientific Computing Department, Science & Technology Facilities Council, Research Complex at Harwell, Didcot, OX11 0FA, United Kingdom. Tel: +44 1235 56 7725; Fax: +44 1235 567720; E-mail: eugene.krissinel@stfc.ac.uk

1 Abstract

New software for fast screening of the Protein Data Bank in 3 dimensions is presented. The software represents a multi-threaded version of the *Gesamt* algorithm, described in a separate publication. The software is implemented as a Qt-powered desktop application *QtGesamt*, dedicated "*Structural Alignment*" task in the new CCP4 web-application "*jsCoFE*" (http://ccp4serv6.rc-harwell.ac.uk/jscofe/), as well as a Restful API for performing remote PDB searches from ordinary PC connected to the internet. The software takes advantage of multi-core architectures, typical for all modern computational platforms, for making fast structural searches in the PDB as well as pairwise and multiple alignments of protein structures in 3d. A combination of parallelization and data reduction techniques made it possible to perform structural queries on modern laptops with speeds that were previously obtainable only from specialized web-servers. A detail performance analysis of *QtGesamt* is given. *QtGesamt* and the corresponding Restful API are available for download from http://ccp4serv7.rc-harwell.ac.uk/gesamt.

2 Keywords:

Structure alignment; Structure superposition; Homology studies;

3 Introduction

Proteins are complex biological polymers, which play central role in the functioning of the cell. The ability of proteins to interact with each other and other biomolecules depends on many factors, particularly the structural complementarity of interacting surfaces. Therefore, comparative analysis of protein structures (essentially the 3D folds of respective polypeptide chains) plays an important role in modern molecular biology. Assuming that similar structures imply similar interactions, detection of structural similarity may help a researcher to identify the biochemical function of a particular protein. Identification of 3-dimensional commonalities is also a routine sub-task in other studies and techniques, such as preparation of models for protein structure solution by molecular replacement.

Usually, comparison of protein structures involves their alignment and superposition. The alignment aims to find geometrically matching parts of compared macromolecular chains, with respect to possibly missing parts and natural dispersion of backbone atoms. The superposition brings matched parts in the closest relative position and orientation, thus allowing for the calculation of a similarity score.

The number of various algorithms for structural alignment and superposition, developed in last 3 decades, is significant (short overviews may be read from, e.g., Refs. [1-3]). The corresponding article in Wikipedia (https://en.wikipedia.org/wiki/Structural_alignment_software) lists 102 software implementations, and this is probably not a complete list. This scale of variety suggests that the problem has not received full solution to date. Indeed, there are no universal definition and measure of structural similarity, which causes the major difference between different algorithms. In addition, structural comparison is NP-complete a problem, therefore, existing algorithms deliver optimal, rather than exact, solutions, and employ different sets of approximations, parameterization and heuristics, by this adding to differences between various methods.

While comparison of several protein structures is relatively quick and efficient, existing algorithms still require a significant

Citation: Krissinel E, Uski V (2017) Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB. J Comp Sci Appl Inform Technol. 2(2):1-7

CPU resource for the routine task of detecting similar structures in the whole Protein Data Bank (the PDB) [4]. For example, structural alignment of a single protein pair, or even multiple alignment of an average-size structure family (10-50 structures) usually takes between 0.01 and few seconds, which is totally acceptable. However, finding structural neighbours in a PDB-size dataset (around 110,000 structures containing some 277,000 chains) requires $\sim 10^5$ more time, or more than a day if done with a personal laptop or desktop computer. This is why database searches are assisted by specialized web-servers, such as SSM [3], DALI [5], CE [5], VAST [6], FATCAT [7] and others. Such servers usually utilize parallel processing on computational clusters in order to deliver results on a sub-minute to hour time scale.

In this communication, we intend to show that the combination of progress in computing hardware, on one side, and advances in algorithms for protein structure alignment, on another side, made it possible to achieve speeds of structural queries that are comparable with what is delivered by specialized servers, on common laptop and desktop computers. Desktop utility *QtGesamt*, built on top of the *GESAMT* (General Efficient Structural Alignment of Macromolecular Targets) algorithm [8], is capable of identifying similar structures in the PDB on sub-minute to an hour time scale. The application has a developed user interface and project bookkeeping, supports visualization of calculated alignments and superpositions with an integrated or external molecular graphics viewers, and may be found to be a convenient alternative to web-servers in many cases, such as working with in-house data archives and when transmission of sensitive data to remote web-servers is not desirable. *QtGesamt* is freely available for download and use on Mac OS X, Linux and Windows platforms. Functionally equivalent application is available as part of new CCP4 online (cloud) services, *jsCoFE*, and also as a Restful API for remote calculation of structural queries. In the rest of the paper, all references, made to *QtGesamt*, are fully applicable to online setups.

4 Basic Methodology Behind Fast Queries

Currently, the PDB contains over 110,000 entries (cf. (http://www.pdb.org), totaling over 26 GB of compressed data. A mere reading and parsing of such amount of data takes about 2 hours with today's most productive PCs (i7 Intel CPUs, Solid-State Disks). The ambition to bring the total time of structural searches in the PDB down to minutes range may be fulfilled with a set of 3 basic elements: a) data reduction and optimized archival b) parallel processing c) optimal choice of targets and alignment seeds (controlled complexity).

4.1 Data Reduction and Optimized Archival

From a whole PDB entry, structural alignment algorithm needs only a subset of atomic coordinates and no metadata, which makes a considerable part of PDB file. E.g., our chosen structural alignment method (*GESAMT*) uses only coordinates of protein's C-alpha atoms. Therefore, *QtGesamt* reduces the PDB archive by selecting C-alpha coordinates and storing them in binary files ("*Gesamt archive*"). Such reduction accelerates similarity searches by, firstly, reducing the amount of data to read and, secondly, eliminating the need for parsing data. Since atomic coordinates in the PDB are presented in fixed format with 3-digit mantissa, it was found beneficial to store them as short integers in order to decrease the file size. In addition, *Gesamt* archive files are compressed as it was found that decompression is faster than reading longer files. The resulting size of *Gesamt* archive is about 1.55 GB, which makes reading at least 15 times faster on comparison with the vanilla PDB archive. The data reduction process usually takes a few hours. In order to cope with ever-growing size of the PDB, a special update mode of data reduction is implemented, in which only new entries are preprocessed and added to the already existing *Gesamt* archive. The update takes only a few minutes every week.

4.2 Parallel Processing

PDB scans are particularly convenient for multi-threading, because structure comparisons are completely independent of each other. In order to fully utilize multi-threading capabilities of modern CPUs, *Gesamt archive* is split in series of files and computations are organized in such a way that parallel threads do not compete for file read. The total number of files should not be too high in order to decrease the overhead associated with file opening. *QtGesamt* runs a configurable number of threads, which depends on hardware used. The number of parallel CPU threads, which can run simultaneously in real time, varies from 2 for 6-8 year old chips up to 16 for today's Intel Haswell (advanced models allow for 20 independent threads). Despite nearly ideal parallelization of PDB scans in *QtGesamt*, the gross acceleration effect is always sub-linear on the number of threads used, see discussion below.

4.3 Optimal Choice of Targets and Alignment Seeds

Data reduction and parallel processing result in 10-15 times faster searches, but the total time still remains in range of hours. Further reduction may be achieved by noting that, due to protein structure diversity, most of protein pairs in the PDB represent dissimilar structures. Ironically, detection of dissimilarity is more laborious a job than alignment of similar structures, therefore, most of CPU resource is spent on comparison of dissimilar structures where results are not interesting. In order to help the situation, *QtGesamt* introduces the minimum parts (similarity levels) R_q and R_t of query and target structure, respectively, which should be possible to align in principle. These parameters are used in two different ways. Firstly, they form a filter for rejection of structure pairs without any attempt to align them: only targets of size N_t, such that

$$N_t^{min} = N_q R_q \;\leq\; N_t \;\leq\; N_t^{max} = N_q / R_t \quad (1)$$

are accepted (in Eq. (1), N_q stands for the size of query structure). Secondly, similarity levels R_q and R_t are used also for the rejection of alignment seeds in the core *GESAMT* algorithm. In its basics [8], *GESAMT* explores a manifold of suitable short fragment clusters, serving as alignment seeds. One can lay origins of these seeds on a $N_t \times N_q$ size matrix shown in Figure 1. It may

Citation: Krissinel E, Uski V (2017) Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB. J Comp Sci Appl Inform Technol. 2(2):1-7

be shown that alignment seeds with C-alpha pairs, found in side triangles of this matrix, cannot reach the minimum required size of $max\left(R_qN_q, R_tN_t\right)$ and, therefore, should not be developed. The combination of filtering and early rejection of unpromising alignment seeds brings further acceleration of PDB queries estimated by a factor of 10-15.

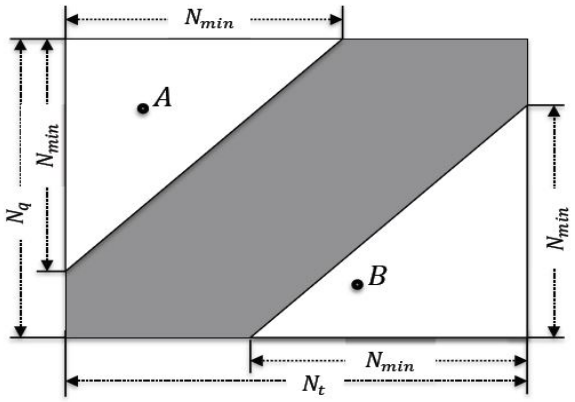

Figure 1 Exclusion of unpromising alignment seeds in QtGesamt. Each point in this matrix, such as A or B, represents a pair of C-alpha atoms f rom query structure of size N_q and target structure of size N_t. The length of alignments that include points A or B from side triangles cannot excee d the minimally required $N_t^{min} = max(N_qR_q, N_tR_t)$ pairs (where R_q and R_t are the minimum similarity levels for query and target structure, respecti vely). Therefore, such alignments are rejected without attempting in orde r to decrease the calculation time. See details in Ref. [8].

5 Implementation

QtGesamt is implemented as a Qt graphical application and is available on all Mac OSX, Linux and Windows platforms (both Qt4 and Qt5 may be used). Algorithmic routines from original *GESAMT* [8] are used as a library and run in threads, rather than processes. Figure 2 shows *QtGesamt*'s graphical interface, which contains a toolbar, project panel and result pages. The interface

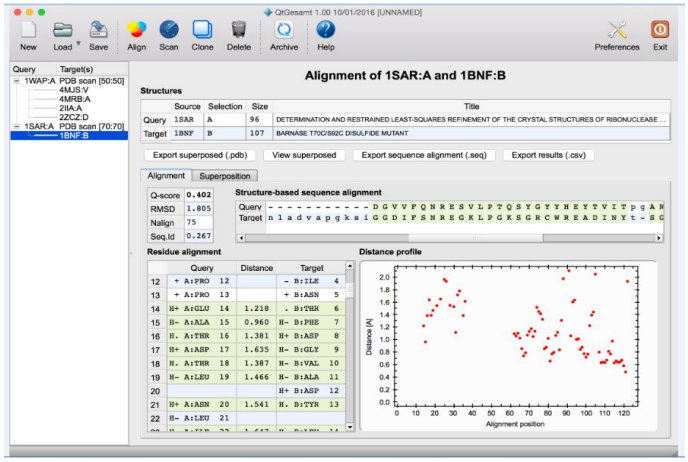

Figure 2 QtGesamt user interface

supports the following main tasks: creation and incremental (i.e. without full re-generation) update of *Gesamt* archive from vanilla PDB archive, alignment of selected structures (both pairwise and multiple), and fast structural searches in *Gesamt* archive. Input structures may be specified by either PDB code or loaded from PDB/PDBx/mmCIF-formatted disk files. On Linux and Mac OSX platforms, *QtGesamt* can work directly with gzipped files. All tasks are automatically stored in the project panel, ordered by query structure. Result pages contain output data, arranged in scrollable tables. Structural superpositions can be visualized using either a built-in (based on JSmol software [9]) or external (CCP4MG [10]) molecular graphics viewers. The built-in viewer is implemented in Javascript and, therefore, may be slow on particular hardware. CCP4MG is implemented with Open GL support and provides excellent user experience. *QtGesamt* interface also makes it possible to save all project or individual task results (alignments, scores, lists of structural hits, superposed structures) in external files, as well as to launch subsequent tasks, such as alignment of selected structural hits from the results of a PDB scan. Full details and functionality of *QtGesamt* interface are available from the built-in program documentation, and also from *QtGesamt* web site. CCP4 [11] distributes a command-prompt version of *GESAMT* software, which is functionally identical to *QtGesamt*.

Table 1 Selected PDB chains and their length in number of aminoacid residues

PDB chain	2AKF: A	6INS: E	1SAR: A	1OCY: A	3ANZ: A	4N4K: A
Size	32	50	96	198	403	497

6 Results and Discussion

6.1 Performance Analysis

Performance of the *GESAMT* algorithm was discussed in Ref. [8]. In particular, it was found that *GESAMT* shows very good discrimination properties (i.e. is both sensitive and selective in discrimination between different protein families and folds), and yields alignments with considerably higher Q-scores than SSM [4] (both aligners are based on the maximization of the Q-score (cf. Ref [4]) and, therefore, can be compared directly). *QtGesamt* inherits all *GESAMT*'s properties; therefore, they will not be investigated here. Instead, we are interested in *QtGesamt*'s suitability for PDB screening.

As shown in Ref. [8], alignment time in *GESAMT* correlates linearly with the product $N_q \times N_t$, and also depends on geometrical properties of aligned structures, showing a variation within almost 2 orders of magnitude for given values of $N_q \times N_t$. The PDB scan time depends also on the number of suitable target structures in the PDB and similarity levels R_q, R_t set in advance. In order to probe the overall time range, we selected randomly 6 PDB structures of different folds and sizes (cf. Table 1), and performed PDB scans at $R_q = 0.7$ and R_t varying from 0 to 0.7 on computational platforms specified in Table 2. Values of $R_{q/t} = 0.7$ are suitable for the identification of close to medium structural neighbours; lower similarity levels are needed for the detection of more remote structures. The results are presented in Figure 3. As seen

Citation: Krissinel E, Uski V (2017) Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB. J Comp Sci Appl Inform Technol. 2(2):1-7

Table 2 Summary of computational platforms tested

Platform Id	Name	Year	CPU	No. of parallel threads	Hard disk
A	Mac Book Pro	2015	Quad-Core Intel i7 2.8 GHz	8	High-speed SSD
B	Mac Book Pro	2009	Intel Mobile Core Duo 2.66 GHz	2	SSD
C	Dell Inspiron 1525	2008	Pentium Dual-Core 1.8 GHz	2	Ordinary HD

from the Figure, calculation times are significantly different for different-size structures and computational platforms. It may be concluded that platforms A and B show very similar per-thread performance, which is about 40% higher than that of platform C. On platform A, the fastest PDB scan (for the smallest-size structure 2AKF: A, 32 residues, $R_t = 0.7$ completed in only 1 second, while the longest one (4N4K:A, $R_t = 0$) took about 50 minutes.

It is interesting to note that the effect of the similarity level R_t on the calculation time depends on structure size. As may be inferred from Figure 4, this is due to the varying number of targets selected by the filtering procedure described above. Indeed, Eq. (1) suggests that for small-size queries, nearly the whole PDB will be selected as a target at $R_t = 0$, while at high R_t, a relatively small subset of targets is explored. However, the larger query structure, the fewer targets in the PDB that can satisfy the upper limit in Eq. (1) at $R_t = 0$, and the number of selected structures ceases to depend on R_t.

Figure 5 presents the average calculation time per alignment T_A on platform A. The decrease of T_A with increase in R_t demonstrates the effect of the rejection of unpromising alignment seeds (cf. Figure 1). The effect is less pronounced in case of larger query structures. This may be understood if one notes that the effect is solely due to the variation of the area of side triangles in Figure 1 with varying R_t. However, the larger N_q, the fewer target structures can have size N_t which is significantly different from N_q and, therefore, the side triangles cease to depend on R_t.

As seen from Figure 4, the total number of accepted target structures may depend significantly on the similarity level . The important question is whether such aggressive filtering has a noticeable impact on the number of relevant targets detected. This question is difficult to answer in general, because "relevance" does not have a clear mathematical definition. In *GESAMT*, relevance is associated with the Q-score [4,8]. It was suggested in Ref. [8] that $Q \geq 0.262$ indicates (probabilistically) similar SCOP families, and $Q \geq 0.1$ indicates similar folds. Figure 6 presents the total numbers of structural hits in PDB scans from Figure 3, which satisfy these conditions. As seen from the Figure, filtering affects noticeably the number of relevant hits only in case of two smaller structures (2AKF:A and 6INS:E), which also give the longest lists of hits. One possible explanation to this fact is that short chains form relatively simple structures, which have higher chances to be found as parts of more complex folds, in which case no particular value of Q-score may be taken as a measure of relevance.

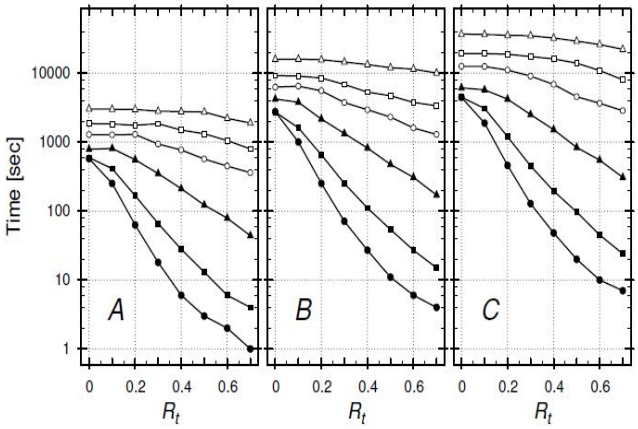

Figure 3 Calculation real times of 6 PDB queries from Table 1. Filled circles: 2AKF:A; filled bars: 6INS:E; filled triangles: 1SAR:A; open circles: 1OCY:A; open bars: 3ANZ:A; open triangles: 4N4K:A. The query similarity level R_q was set to 0.7 and the target similarity level R_t varies along horizontal axes. Plots A, B, and C correspond to computational platforms specified in Table 2. The maximal number of parallel threads (cf. Table 2) was used.

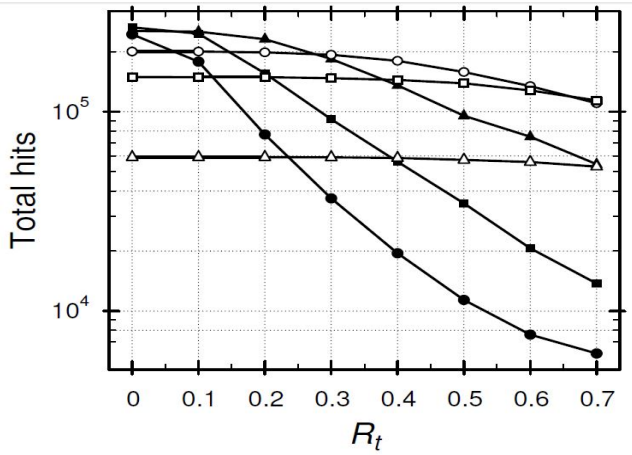

Figure 4 Total number of structural hits for PDB scans presented in Figure 3. The curve labeling is the same as in Figure 3.

Finally, consider the effect of multi-threading on the scan time. As seen from Figure 7, scan time depends sub-linearly on the number of used threads n_t, reaching the lowest value at the maximum number of threads that can be run simultaneously on given computational platform. This is a well-known feature of multi-threaded systems, which, in general, is due to two main factors: 1) not a full parallelization of the computational code, which causes threads to compete for memory and disk access, and 2) growing overhead of intra-CPU communication, competition and

Citation: Krissinel E, Uski V (2017) Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB. J Comp Sci Appl Inform Technol. 2(2):1-7

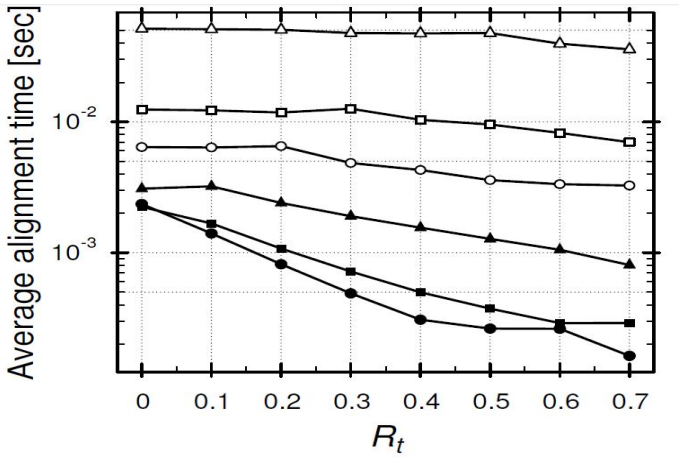

Figure 5 Average alignment time as a function of target similarity level R_t for platform A (cf. Table 2). The curve labeling is the same as in Figure 3.

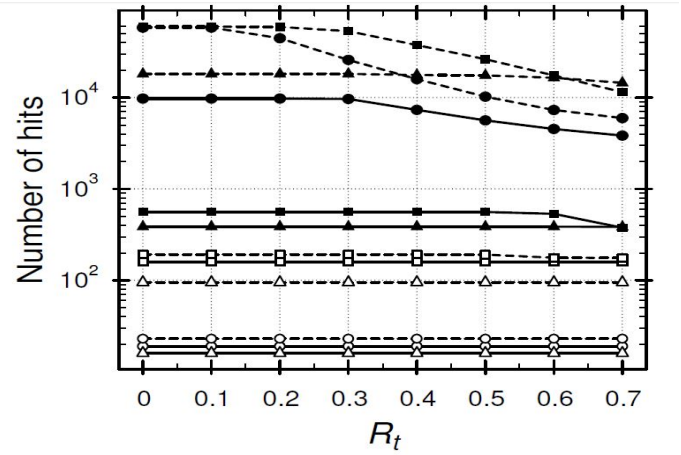

Figure 6 Total number of structural hits with Q-score higher than 0.262 (solid lines), and higher than 0.1 (dashed lines), as a function of target similarity level R_t. The curve labeling is the same as in Figure 3.

switchover between the growing numbers of threads. First factor does not apply in case of *QtGesamt*, where threads are completely separated by design. This was checked by measuring execution times of variable number of single-threaded *GESAMT* processes, which gave the same per-thread yields as in Figure 7. One can also see from the Figure that the maximum speed-up amounts to 3.5 at $n_t = 8$, of which 2.8 is achieved at $n_t = 4$. This is likely to be associated with the fact that system's A CPU is made of 4 hyper-threaded cores, while hyper-threading is not efficient at high CPU loads. It may seem to be a good idea to use no more than 4 threads, gaining 80% of QtGesamt productivity but leaving 50% of CPU free for other tasks. However, due to factor (2), running other tasks will slow down all the threads and, in reality, little benefit can be obtained.

6.2 *GESAMT* software in CCP4 Cloud

QtGesamt represents a convenient and robust tool for protein structure alignment and database searches in 3 dimensions, which also supports project structure for bookkeeping user's

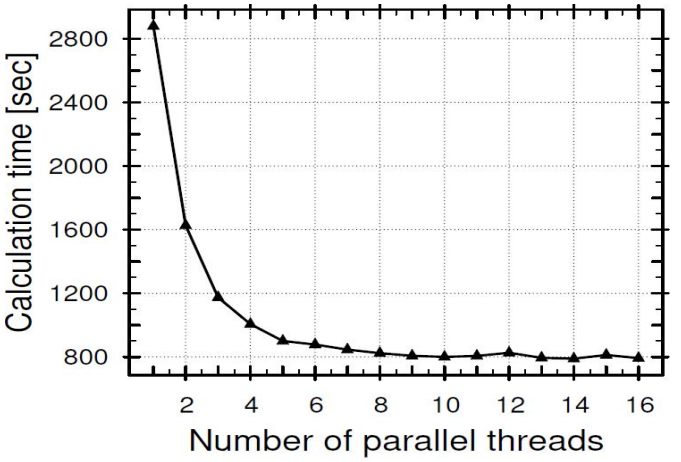

Figure 7 Calculation real times of PDB query 1SAR: A at N_q=0.7 and N_t=0 obtained on system A for different number of computational thr eads.

work. However, very often, structure similarity studies are performed as part of a larger task, for example, the choice and construction of models for molecular replacement in protein crystallography [12]. In such cases, structure alignment software is seamlessly incorporated in computational pipelines, yet, the need for direct manipulation of the alignment tool as part of a larger structure solution framework is still there. In order to address such situations, *Gesamt* was included in CCP4 Graphical Interface [11] and the new CCP4 online service for structure solution, *jsCoFE* (javascript-based Cloud Front End). The corresponding task interface in jsCoFE is shown in Figure 8, and Figure 9 exemplifies a representative part of the output. *jsCoFE* provides own facilities

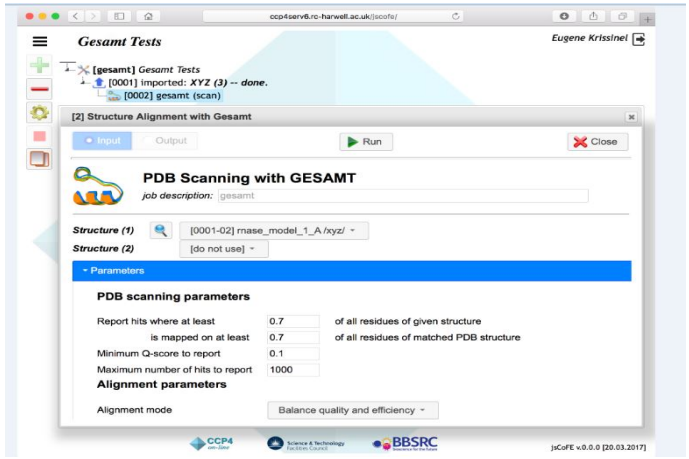

Figure 8 Gesamt user interface in jsCoFE, the new CCP4 online service for structure solution.

for the management of computational jobs and project structure, where *Gesamt* jobs are presented in combination with other crystallographic tasks, communicating by means of predefined data objects behind the scene (in the cloud). The corresponding data flows are reflected in a tree-like project structure, where nodes represent particular tasks performed.

Online use of *Gesamt* software does not provide any substantial

Citation: Krissinel E, Uski V (2017) Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB. J Comp Sci Appl Inform Technol. 2(2):1-7

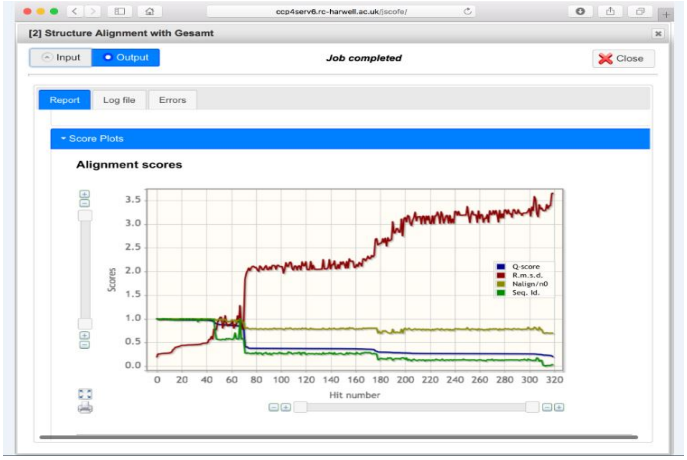

Figure 9 An excerpt from Gesamt output in jsCoFE, showing the relation between various alignment scores obtained from the PDB scan for structure model highly similar to PDB entry 1SAR:A.

speed-up on comparison with modern desktops, but it benefits from the integration with other crystallographic tasks and seamless support of maintained PDB and *Gesamt* archives on CCP4 Cloud servers. Being a rather universal tool, *Gesamt* has many uses outside crystallographic context, for example, in bioinformatics studies involving the assessment of structural similarity. In such cases, researchers can also benefit from computational and database support in CCP4 Cloud by using *Gesamt* software through a dedicated Restful API. In this mode of operation, Restful calls are used instead of direct invocation of *Gesamt*. The API packs all necessary data and instructions and sends them to a dedicated gateway in CCP4 Cloud. After job completion, the results can be retrieved by a successive Restful call, which delivers a package containing list of structural hits found, as well as report similar to one used in *jsCoFE*. Instructions on the installation and use of Gesamt's restful API are found in *QtGesamt* main page: http://ccp4serv7.rc-harwell.ac.uk/gesamt.

6.3 Further Directions

We represent that Gesamt software is one of fastest methods for macromolecular structure alignment and superposition in 3d. It remains somewhat slower than, e.g. SSM [3,8], which was found to be the fastest one in Ref. [13]. However, use of Gesamt is fully justified by its high protein fold discrimination power and independence of protein's secondary structure [8]. But even with the speed achieved, simple estimates suggest that cross-alignment of all structures in the PDB will take about 10 CPU-years and require some 1TB disk space. These numbers often deter researchers from the full-scale structural analysis of the PDB. Such analysis is needed, for example, in order to improve protein structure classification, important for the inference on protein's biochemical function and overall role in the biochemistry of the cell. In order to facilitate the analysis, reduced "representative", non-redundant, datasets are often used instead of full archive. A similar situation is observed also in the preparation of models for molecular replacement in protein crystallography, where, typically, limited non-redundant subsets of structures are used. While the removal of redundant structures is beneficial in many cases, the success of molecular replacement is known to be highly dependent on structural variations within 1-1.5 angstroms, which is typically far beyond the usual redundancy thresholds. Having a complete matrix of cross alignments in the PDB would also accelerate structural searches by a more appropriate choice of promising structure pairs. Therefore, we see the database support for *Gesamt*, in terms of maintainable full cross-alignment matrix, as a priority direction for further research and development. Any practical use of such matrix is feasible only in the cloud-computing framework, which becomes more and more attractive in many instances, particularly in macromolecular crystallography and structural bioinformatics.

7 Conclusion

We described the application of GESAMT algorithm for structural searches in the Protein Data Bank, which required parallelization and appropriate preparation of structural data. The developed software utilizes multi-threading capabilities of modern computational platforms, which makes it possible to perform full PDB scans in real time, using common laptop and desktop PCs. In this communication, we demonstrated these capabilities on a range of PDB structures and computational platforms, which were chosen to cover most of practical cases. This study should not be considered as a comprehensive one, and performance variations will be most definitely observed from case to case. As a general conclusion, fast PDB scans are quite feasible on modern systems of personal computing, such as platform A in Table 2, while older systems (B, C and similar) may appear slower than fully comfortable. With rapid development of computing hardware, especially with the growing number of cores per CPU, we expect that the usefulness of *QtGesamt* software as an alternative to specialized web-servers will only increase.

On the other side, it is acknowledged that the use of *QtGesamt* requires local management and regular update of considerable resources: the PDB archive (nearly 26 GB of data) and *Gesamt* archive derived from it (close to 1.6 GB). Therefore, the need in online (cloud) setups does not completely disappear. We address this problem with the development of Gesamt task for PDB searches and structural alignment in new CCP4 Cloud resource *jsCoFE*, and by providing Restful API to perform remote Gesamt tasks from local scripts.

8 Acknowledgement

This work was supported by research grant BB/L007037/1 "CCP4 Grant Renewal 2014-2019: Question-driven crystallographic data collection and advanced structure solution" from the Biotechnology and Biological Sciences Research Council (BBSRC) UK. The authors are also thankful to Collaborative Computational Project Number 4 in Protein Crystallography, UK (CCP4) for supporting *GESAMT* developments, outstanding maintenance and distribution effort involved, as well as for providing many opportunities for associated education and dissemination activities. The authors would like to thank Dr. Jens Thomas, University of Liverpool, and Dr. Ronan Keegan, CCP4, for stimulating discussions.

Citation: Krissinel E, Uski V (2017) Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB. J Comp Sci Appl Inform Technol. 2(2):1-7

9 References

1. Sippl MJ, Wiederstein M. Detection of spatial correlations in protein structures and molecular complexes. Structure. 2012;20(4):718-728. doi:10.1016/j.str.2012.01.024

2. Theobald DL, Steindel PA. Optimal simultaneous superpositioning of multiple structures with missing data. Bioinformatics. 2012;28(15):1972-1979. doi: 10.1093/bioinformatics/bts243

3. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004; 60(Pt 12 Pt 1):2256-2268. doi:10.1107/S0907444904026460

4. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Protein Eng. 1998;11(9):739-747. doi: 10.1093/protein/11.9.739

5. Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. Nucleic Acids Res. 2010;38(Web Server issue):W545-549. doi: 10.1093/nar/gkq366

6. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol. 1996;6(3):377-385. doi:10.1016/S0959-440X(96)80058-3

7. Ye Y, Godzik A. FATCAT: a web server for flexible structure comparison and structure similarity searching. Nucleic Acids Res. 2004;32(Web Server issue):W582-W585. doi: 10.1093/bioinformatics/btg1086

8. Krissinel E. Enhanced fold recognition using efficient short fragment clustering. J Mol Biochem. 2012;1(2):76-85.

9. Hanson RM, Prilusky J, Renjian Z, Nakane T, Sussman JL. JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. Israel Journal of Chemistry. 2013;53(3-4):207-216. doi:10.1002/ijch.201300024

10. McNicholas S, Potterton E, Wilson KS, Noble ME. Presenting your structures: the CCP4mg molecular-graphics software. Acta Crystallogr D Biol Crystallogr. 2011;67(Pt 4):386-394. doi:10.1107/S0907444911007281

11. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. Acta Crys. 2011;D67:235-242. doi:10.1107/S0907444910045749

12. Read RJ. Pushing the boundaries of molecular replacement with maximum likelihood. Acta Crystallogr D Biol Crystallogr. 2001;57(Pt 10):1373-1382.

13. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein alignment methods: scoring by geometric measures. J Mol Biol. 2005;46(4):1173-1188.

Citation: Krissinel E, Uski V (2017) Desktop and Web-based GESAMT Software for Fast and Accurate Structural Queries in the PDB. J Comp Sci Appl Inform Technol. 2(2):1-7