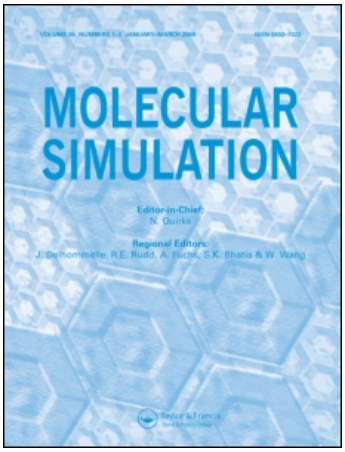

Molecular Simulation

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

A grid-enabled algorithm yields figure-eight molecular knot

Fotis E. Psomopoulos[a]; Pericles A. Mitkas[a]; Christos S. Krinas[b]; Ioannis N. Demetropoulos[c]
[a] Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece [b] Department of Molecular Biology and Genetics, Democritus University of Thrace, University Campus, Alexandroupolis, Greece [c] Department of Engineering Informatics and Telecommunications, University of Western Macedonia, Kozani, Greece

To cite this Article Psomopoulos, Fotis E. , Mitkas, Pericles A. , Krinas, Christos S. and Demetropoulos, Ioannis N.(2009) 'A grid-enabled algorithm yields figure-eight molecular knot', Molecular Simulation, 35: 9, 725 — 736
To link to this Article: DOI: 10.1080/08927020902833103
URL: http://dx.doi.org/10.1080/08927020902833103

Molecular Simulation
Vol. 35, No. 9, August 2009, 725–736

Taylor & Francis
Taylor & Francis Group

A grid-enabled algorithm yields figure-eight molecular knot

Fotis E. Psomopoulos[a1], Pericles A. Mitkas[a2], Christos S. Krinas[b3] and Ioannis N. Demetropoulos[c]*

[a]*Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki 54 124, Greece;* [b]*Department of Molecular Biology and Genetics, Democritus University of Thrace, University Campus, Alexandroupolis 68 100, Greece;* [c]*Department of Engineering Informatics and Telecommunications, University of Western Macedonia, Kozani 50 100, Greece*

(*Received 2 December 2008; final version received 18 February 2009*)

The recently proposed general molecular knotting algorithm and its associated package, MolKnot, introduce programming into certain sections of stereochemistry. This work reports the G-MolKnot procedure that was deployed over the grid infrastructure; it applies a divide-and-conquer approach to the problem by splitting the initial search space into multiple independent processes and, combining the results at the end, yields significant improvements with regards to the overall efficiency. The algorithm successfully detected the smallest ever reported alkane configured to an open-knotted shape with four crossings.

Keywords: knot theory; stereochemistry; grid computing; data decomposition; figure-eight molecular knot

1. Introduction

A particular spatial entanglement of a molecule is a mechanically interlocked molecular architecture that is analogous to a macroscopic knot (Figure 1). Molecular knots are also referred to by some chemists as 'knotanes'. The term knotane was coined by Lukin and Vögtle [1] by analogy with rotaxane and catenane. Knotanes are increasingly found in nature (knotted proteins [2,3] and DNA [4]), while synthesised molecular knots [5,6] along with catenanes and rotaxanes have been proposed as parts of potential molecular machines [7–9] or scaffolds [10] for drug carriers. Nowadays, molecular modelling, such as computational materials and drug discovery, are among the largest consumers of CPU power (probably excluding military applications second only to weather prediction) [11].

Existing molecular knot construction methods [12–16] are considered inadequate for short molecular chains. Recently, we have proposed a *de novo* computer-aided systematic production of small molecular knots with three or more crossings and unlinked ends [17]. The generalised molecular knotting algorithm (GMK) was implemented in the public domain code MolKnot [18]. The GMK algorithm narrows the knotted conformer search to a subdomain of the total conformational space (TCS), which is most likely to contain knotted conformers according to the topological theory of open polygonal knots [19,20]. The algorithm itself is amenable to domain decomposition, so the program has been modified in order to run in a computational grid.

Most of the modelling in physics and chemistry is of the type of quantitative structure–activity relationship (QSAR) [21], which describes how a chemical structure is quantitatively correlated with a well-defined property, such as biological activity, chemical reactivity or physical property. The general mathematical form of the QSAR is as follows (Equation (1)):

$$\text{Activity} = f(\text{physicochemical properties and/or structural properties}). \quad (1)$$

The allocation of a value to a molecular three-dimensional structure's property, such as similarity to a template, solubility, lipophilicity, stability (expressed as an energy function minimum), dipole moment etc, is distinct for each property. For instance: (1) a QSAR connects the dynamic viscosity (η) of dilute solutions of the triglycerides to the viscosity of the solvent (η_0), the triglycerides' concentration (C) and the structural characteristics, such as the length of the carbon chains (CN) and the number of double bonds (DB) via the simple empirical equation $\ln(\eta) = k_0 + k_1\ln(\eta_0) + k_2\text{CN} + k_3\text{DB} + k_4\text{C}$ [22]. (2) A QSAR on enzymic activity could be an inequality constraint. Root mean square distances between the heavy atoms of the side chains of certain catalytic triad residues of an enzyme should be less than 2 Å [23], so an enzymomimetic compound [24] could fulfil the spatio-temporal criterion [25]. In maths, the knottiness of a polygonal line is related to certain relationships between its linear segments [19,20]. The molecular knottiness is the conformers' subspace (total knotted conformer space (TKCS) – of a compound). TKCS is a function of the macromolecule's length and the number of its linear segments. Hence, the related topological indicator is knottiness $<f$ (macromolecules

*Corresponding author. Email: idimitr@uowm.gr

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020902833103
http://www.informaworld.com

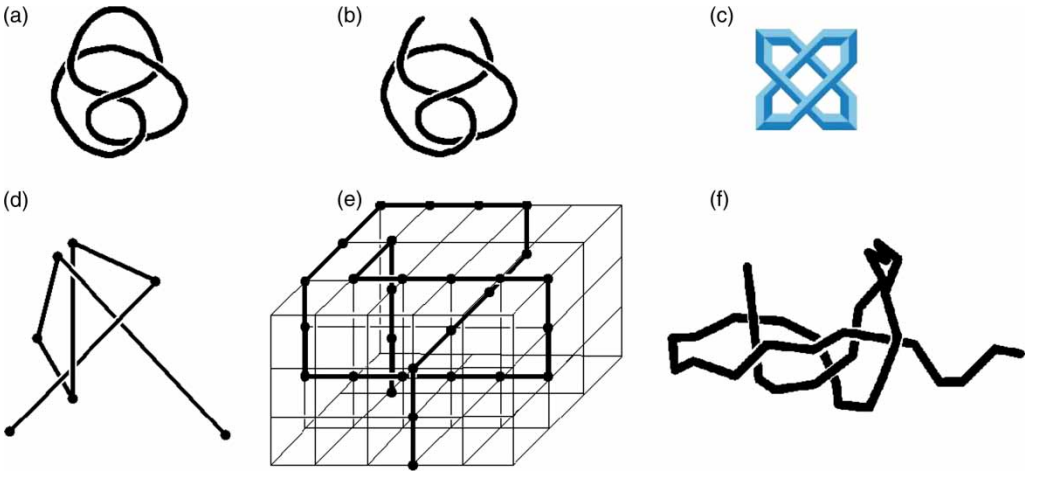

Figure 1. Figure-eight knots in various settings. (a) A rope shaped as 41 knot (figure-eight knot), (b) a rope with open ends, (c) a 41 knot using two closed hexagons (also known as King Solomon's knot), (d) a polygonal chain with seven vertices in open knot 41 formation, (e) open figure-eight knot constructed with 29 vertices (which comprise 11 linear segments with lengths as follows: 2, 3, 1, 3, 2, 2, 5, 2, 4, 1 and 3), and (f) an alkane as an open figure-eight knot.

length, number of linear segments) [17]. Furthermore, it has been shown [27] that knottiness induces bond elongation and valence angle distortion so a possible QSAR is reactivity $= f$ (knottiness, bond_elongation, valence_angle distortion).

Since the problem of finding the property's value on the subdomains (different molecules or conformers) is independent to each other, the QSAR construct is ideally suited as a grid-computing application.

The structure of this paper is as follows. Section 2 describes the implementation of the gridification of the MolKnot program, Section 3 presents a case study from the homologous hydrocarbon series C_nH_{2n+2} ($C_{36}H_{74}$) with emphasis to 4_1 knots. Finally, the main findings of the work are summarised in Section 4.

2. Computational methods

2.1 The GMK algorithm

An interesting aspect of linear alkane molecules is that they have the ability to fold, thus creating knots in three-dimensional space. The molecule chain can be modelled as a polygonal line, allowing for an algorithmic approach to the problem of detecting the specific segment that can produce a knot. The generalised molecular knot algorithm is a four-step computational approach that targets open knot formations. This original procedure is as follows:

(1) *Conformer generation step:* Molecular conformations with inherent knot topology are created by applying the theory of open polygonal knots at dihedral space,
(2) *Math filter step:* Each structure is evaluated for knottiness by calculating the Alexander polynomial $\Delta(t)$ [17,26,27],
(3) *Molecular mechanics (MM) step:* The confirmed knots are subjected to constraint geometry optimisation using classical MM force field equations, and are then re-evaluated for knottiness, and
(4) *Quantum physics step:* The resultant structures are subjected to unconstrained geometry optimisation at a semi-empirical quantum level (AM1).

The flow chart of the GMK algorithm, showing the main conceptual steps in the process, can be seen in Figure 2. This chart does not include the quantum step since that only serves as a shape optimiser [17,27] – knottiness is generally preserved. Moreover, the first three steps of the algorithm are the most computationally expensive. Hereby, we define lengths at the dihedral space of linear molecules. The polyethylene molecule C_nH_{2n+2} is modelled as a finite number of straight line segments that are linked together with suitable turns. Line segments are represented as contiguous 180° CCCC torsion angles, and the turns are $-(CH_2)_5-$ fragments with two contiguous torsion angles (Figure 3(a), (b)). In practice, this means that the angle defined by three consecutive atoms C_i, C_{i+1} and C_{i+2} is rigid (i.e. is not permitted to change more than $\pm 5°$), whereas the dihedral angle defined by four consecutive atoms C_i, C_{i+1}, C_{i+2} and C_{i+3} is flexible, by allowing the rotation along the axis defined by C_{i+1} and C_{i+2}. This is the definition of 'dihedral angle' between four points.

Without getting into detail, each generated molecular conformation must pass through two consecutive filters. First, a mathematical filter rejects any formation that is not knotted (i.e. the formation must adhere to a knot topology, as shown in Figure 3(c)). Second, a MM filter allows only the mathematical knots that are physically accepted

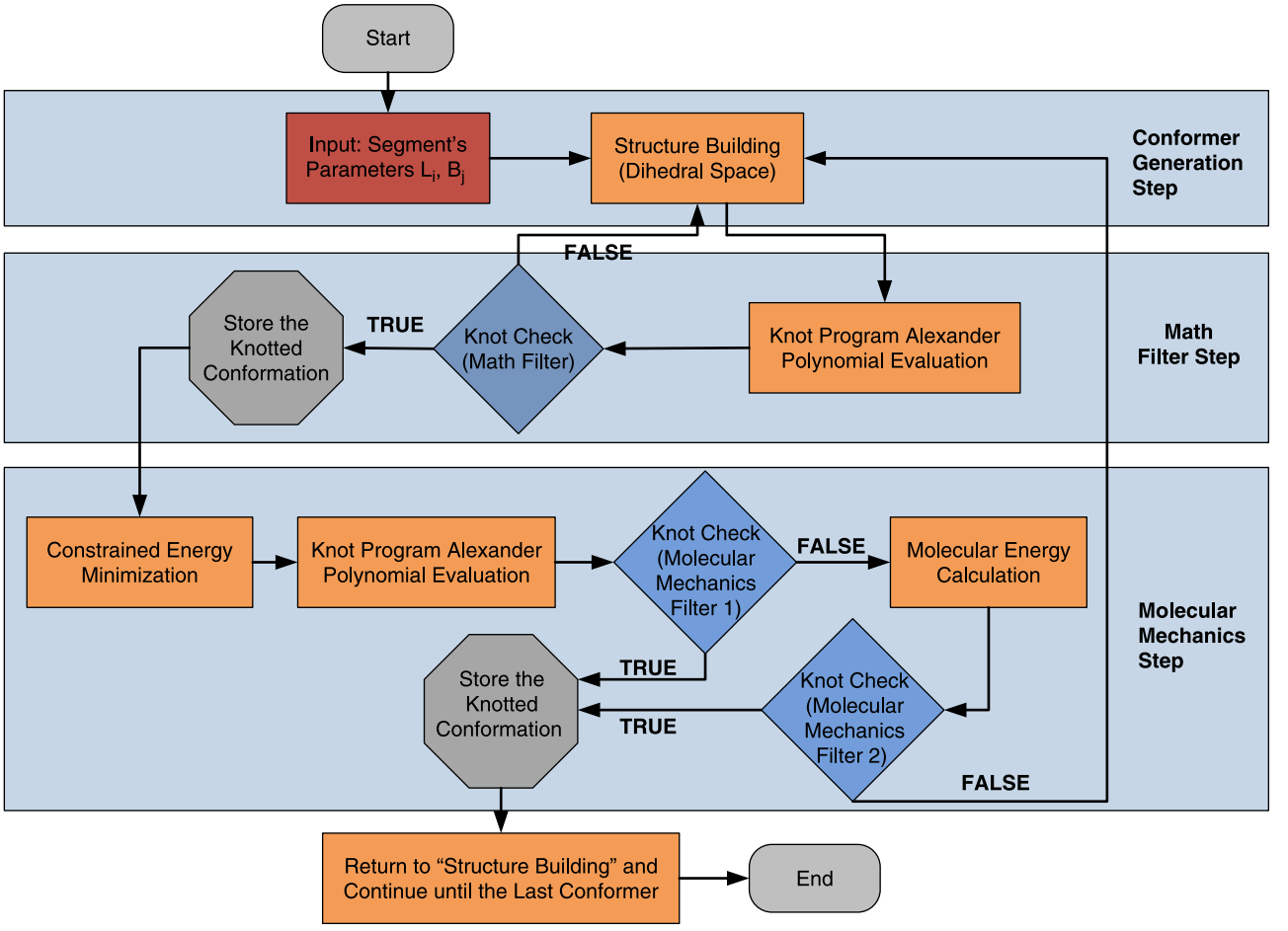

Figure 2. Logical diagram of the GMK algorithm. The conformers are saved as open polygonal knotted graphs and as MMs' optimised structures. This figure clearly depicts the three initial steps as described in Section 2.1. MM filter 1 decides based on the Alexander polynomial evaluation for a geometry optimised using the MM description. MM filter 2 decides based on the energy value assigned to the conformer under examination. A high energy value compared to the energy of the extended $C_{36}H_{72}$ hydrocarbon molecule implies that the conformer is entangled. The energy criterion is especially discriminating for tight knots.

by means of a multi-spring model (i.e. the molecule is described as a multi-spring device). The algorithm has been shown to be quite efficient for short chain lengths, where all other methods fail [27]. The build-up of a 4_1 knot at molecular level is illustrated in Figure 3, while the sequence of twists that lead to a figure-eight knot, realised in a rope, is presented in Figure 4.

The fraction of the knotted conformers (TKCS) over all permitted possibilities for the TCS is a measure of the Delbrücks [29] knotting probability. The TKCS can be evaluated using Equation (2), where *nc* is the number of carbons in the molecular chain and *k* is the number of segments in the polygonal line model.

As an example, the TCS size of $C_{36}H_{74}$ is enormous (1.5×10^{17}), while TKCS is seven orders of magnitude smaller (5.6×10^{10}). By removing segment lengths that are less likely to produce tight knots, the TKCS can be further reduced to 5.5×10^{8}. In [27], the TCS size of $C_{30}H_{62}$ is stated to be 9.9×10^{14}, while TKCS is reported to be seven orders of magnitude smaller (16, 656, 192).

$$\mathrm{TKCS}(nc,k) = \sum_{k=5}^{\lfloor((nc-1)/3)\rfloor} \left(\frac{(nc-2k-2)!}{(k-1)!(nc-3k-1)!} \times 6^{k-1} \right). \tag{2}$$

It is obvious that the process can be very time consuming, as the size of the molecule in question increases. On the other hand, the sequential approach to the problem may not be optimal, due to the fact that each generated conformation could potentially go through both phases, although evaluations of molecular conformations are independent. In fact, the conformations can be regarded as the 'data' of the process, while the rest of the program can be viewed as a single function, thus characterising the GMK algorithm as a data decomposition process. Therefore, as an embarrassingly parallel process, GMK can only gain through deployment over a grid environment.

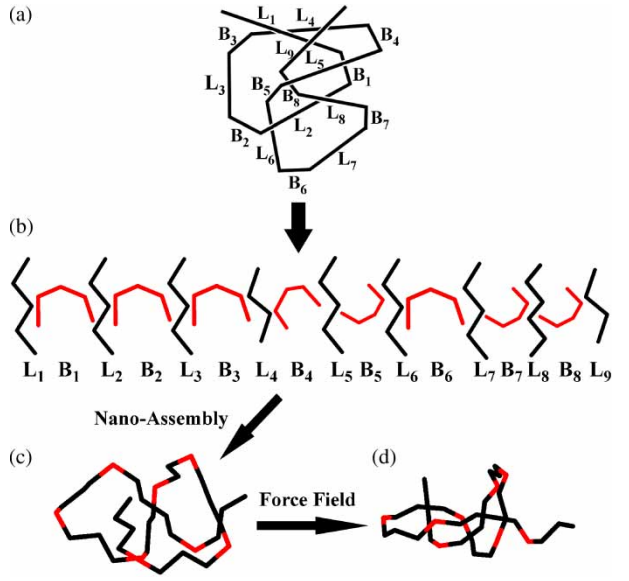

Figure 3. A graphical presentation of the GMK algorithm as applied to $C_{36}H_{74}$ molecule for $k = 9$. (a) The model with nine straight segments L_i and eight turning points B_j. (b) Molecular segments used for the assembly of the initial molecular geometry (L_i, black colour), (B_j, red colour). (c) The initial building sequence $\Phi_1,\Phi_2,\ldots,\Phi_{33}$, where $\Phi_1,\Phi_2,\ldots,\Phi_{33}$ are CCCC torsion angles in the carbon backbone of the $C_{36}H_{74}$ molecule. (d) Tube representation of the outcome of the geometry optimisation procedure of the $C_{36}H_{74}$ hydrocarbon; the sequence of CCCC backbone torsion angles. Appendix A contains the specifics of steps (b), (c) and (d), which may allow the interested reader to reproduce the transformations. The conformer maintained its entanglement at B3LYP/631G** optimisation level. For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.

2.2 *Grid and MolKnot*

Although the term 'grid' has been conflated, at least in popular perception, to embrace anything from advanced networking to artificial intelligence, grid computing has emerged as an important new field in computer science. Foster and Kesselman [30] gives the following definition of the grid:

> a computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to high-end computational capabilities.

In practice, a grid environment can be perceived as a virtual computing architecture that provides the ability to perform higher throughput computing by taking advantage of many computers geographically dispersed and connected by a network. Currently, there exist several major grid middleware providers, such as EGEE [31], UNICORE [32], Globus [33] (Europe, USA) and Vega [34] (CNGrid). However, the differences in software are not reflected on the actual hardware infrastructure. In fact, there is an ongoing effort of unifying the grid infrastructures worldwide. The first step toward this direction was the interoperability of the European and USA middlewares, namely EGEE, UNICORE and Globus. There are several projects that aim to integrate the European infrastructure with others, such as EUChinaGrid (which provides specific support actions for the interoperability of EGEE and CNGrid), and the EUIndiaGrid project (which will make available a common infrastructure to the European and Indian scientific community). Due to this interoperability initiative, the G-MolKnot application is expected to run without any problems in other grid infrastructures, but this has yet to be tested.

G-MolKnot was developed based on the MolKnot algorithm, aiming to create an implementation that would take full advantage of a grid environment. Due to the embarrassingly parallel nature of the algorithm (as will be discussed later), the implementation is not based on the MPI (Message Passing Interface) Library, thus enhancing the portability of the application. The design paradigm falls into the parameter sweep category; the conformational range is divided into a number of disjoint subranges and each subspace is assigned to a different process, which performs the MolKnot procedure through all steps. Finally, the results are merged together, making the whole process transparent to the end-user. Although the divide-and-conquer approach has been proved efficient several times in parameter sweeping problems [35], it stands to reason that a decomposition of the MolKnot procedure into multiple independent processes will allow for a significant gain in execution time.

2.3 *GMK complexity analysis*

As shown in the previous section, the GMK algorithm can be decomposed into three independent steps. An initial approximation to the overall complexity of the algorithm is the following: given a conformation space V of a molecule with nc number of carbons, the number of generated molecular formations will be $N = |V|$ (N takes

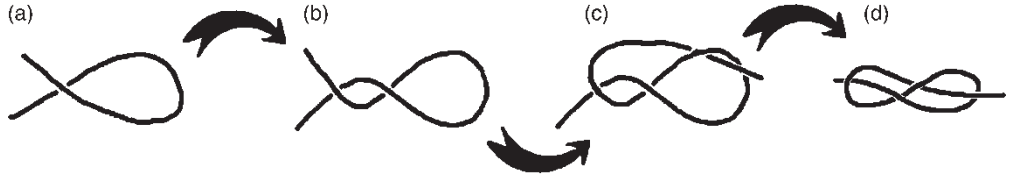

Figure 4. Figure-of-eight or Flemish knot. The rhyme to forming the knot is 'Twist it once, twist it twice; pass it through and make it nice' [28].

the value of TKCS). In the worst case scenario, where all generated formations are indeed actual knots, the sequential approach will have to make N iterations through the first three phases. Without loss of generality, the complexity of the first phase can be assumed to be equal to $O(N \times nc)$, the complexity of the second phase equal to $O(N \times nc^2)$ and the complexity of the third phase equal to $O(N \times nc^3)$. Hence, the sequential complexity would be approximately equal to $O(\sum_{k=1}^{3} N \times nc^k)$. Using the G-MolKnot approach, the same space V will be decomposed to p disjoint subspaces, each of which will be processed independently by one of the p requested CPU units. Thus, the complexity becomes approximately equal to $O(\sum_{k=1}^{3}(N \times nc^k)/p)$. In practice however, the number of formations that successfully pass through each filter decrease exponentially with each phase. That is, if V_i, $i = 1, 2, 3$, is the conformation space at the end of phase i, with $V \supseteq V_1 \supseteq V_2 \supseteq V_3$, then $|V| \gg |V_1| \gg |V_2| \gg |V_3|$, thus leading to a significant reduction in complexity, as shown in the experiments in the next section.

Equation (3) shows the complexity of the first three steps of GMK for a single conformation as a function of nc.

$$O(nc) = \underbrace{z \times O(nc)}_{\text{Molecule Generation}} + \underbrace{x \times O(nc^{2.376})}_{\text{Math Filter}} + \underbrace{\left[y_1 \times O(nc^2) + y_2 \times O(nc^2) + y_3 \times O(nc^3) \right]}_{\text{Molecular Mechanics Filter}}. \quad (3)$$

Experimentally, z, x, y_1, y_2 and y_3 take the following values: $z = x = 1$; $y_1 \approx 900$; $y_2 \approx 800$; and $y_3 \approx 1200$. The complexity of the second term is given according to Coppersmith and Winograd [36].

In a real-world example, the number of conformations N is a significant factor in the overall complexity of the problem, transforming Equation (3) into the following:

$$O(nc, N) = \ln^{2/3}(N) \times \left[\underbrace{z \times O(nc)}_{\text{Molecule Generation}} + \underbrace{x \times O(nc^{2.376})}_{\text{Math Filter}} + \underbrace{\left[y_1 \times O(nc^2) + y_2 \times O(nc^2) + y_3 \times O(nc^3) \right]}_{\text{Molecular Mechanics Filter}} \right]. \quad (4)$$

$$O(nc, N, a, b) = \ln^{2/3}(N) \times \left[z \times O(nc) + x \times O(nc^{2.376}) + \left[\frac{a}{N} \times \left(1 - \frac{b}{a} \right) \times (y_1 \times O(nc^2) + y_2 \times O(nc^2) + y_3 \times O(nc^3)) \right] + \left[\frac{a}{N} \times \frac{b}{a} \times (y'_1 \times O(nc^2) + y'_2 \times O(nc^2) + y'_3 \times O(nc^3)) \right] \right] \quad (5)$$

Finally, a thorough complexity analysis will require a closer look at the design of the MM filter, which comprises three serial iterative procedures. In the worst case scenario, a single molecular conformation that would fail the filter will have to exhaust all available iterations, as opposed to a knotted conformation, which will pass through the filter using a small number of iterations.

If we define:

- a as the number of conformations that pass through the mathematical filter, and
- b as the number of conformations that pass the MM filter, the complexity of the GMK algorithm (for phases 1–3) is presented in Equation (5), where $z = x = 1$, $(a/N) \approx 10^{-3}$, $(b/a) \approx 10^{-2}$, $y_1 \approx 1900$, $y_2 \approx 1800$, $y_3 \approx 650$, $y'_1 \approx 300$, $y'_2 \approx 320$ and $y'_3 \approx 1900$ (experimental values).

In this case, the MM filter complexity has been split into two parts. The first expresses the case when a conformation is rejected by the filter, which applies to the majority of the conformers, and it is usually captured by the first steps. The second part expresses the case when a conformation is accepted by the filter, which means that has successfully passed all three steps. However, in this case each step is passed relatively quickly, and this is reflected in the values of y'_1, y'_2 and y'_3. Overall, it can be seen that the worst case scenario is $O(nc, N) = \ln^{2/3}(N) \times O(nc^3)$. The coefficients in the analytical form of the complexity suggest that the actual complexity is lower and closer to 2 than 3. Moreover, the complexity can be further reduced by optimising the steps that constitute the MM filter, so as to reduce the coefficient of the $O(nc^3)$. By splitting the data (i.e. the conformations) into multiple processes p, the new complexity of the parallel procedure will be

$$O(nc, N, p) = \frac{\ln^{2/3}(N) \times O(nc^3)}{p} \quad (6)$$

which clearly indicates the validity of porting the GMK algorithm to a grid environment. This analysis will be an important feedback for a better tuning of the program, by increasing the ratio of y_1, y'_1, y_2, y'_2 against y_3, y'_3. In this way future runs will perform better.

3. Results and discussion

3.1 Computational experiment set-up

For the experiment, a subspace of Equation (2) will be used (SubKCS(nc, k)), by allowing only a single value of k. Moreover, instead of all six available values of bends, only four were allowed. Therefore, Equation (2) is transformed as follows:

$$\text{SubKCS}(nc, k) = \left(\frac{(nc - 2k - 2)!}{(k-1)!(nc - 3k - 1)!} \times 4^{k-1} \right). \quad (7)$$

The experiment design focused on the $C_{36}H_{74}$ molecule with $nc = 36$ and the polygonal line comprising of $k = 9$ segments (as discussed in Section 2). These parameters lead to a TCS of approximately 3.5×10^9. However, a further reduction in the search space was deemed necessary, so that the acceptable segment lengths of the polygonal line would be limited to 1, 2, 3 and 4. The size of this new subspace is 553,844,736 (order of 0.5×10^9), which is an order of magnitude larger than any other in literature so far [17,27].This means that the corresponding computational power required for the experiment is also high [20]. For this reason, the 'gridification' of the process is justified.

The experiment allows for a fairly complete search of the conformational space of the molecular chain with 36 carbon atoms (dihedral length 33). A longer molecular chain provides the ability to search for knotted conformations of greater complexity than those studied in existing literature. However, even this length is restrictive for higher complexities (such as five crossings or more). Despite this fact, this paper focuses to search for knots with four crossings, which was not possible with previously studied lengths. In this section, the results of the G-MolKnot application are presented, in order to evaluate its effectiveness and efficiency. There is also a discussion on the correlation of the conformational space, which TKCS searched with the number of produced knotted conformers at each filter and their complexity (i.e. knots with three or four crossings).

3.2 Optimisation of run-time parameters for grid execution

The analysis of the grid parameters is performed in order to define the optimal parameters for the final run. On one hand, the major issue in the gridification of the process is to define the number of subspaces that will be assigned to the independent processes on the grid with regards to time efficiency. On the other hand, the effect of the heterogeneity of grid resources on the application should also be studied.

The first step in the optimisation process is to verify in practice that the domain decomposition, as presented in Section 2, is justified as an optimisation process for the specific problem. To this aim, a smaller dataset that comprised a 27 carbon atoms alkane and only part of its conformational space was used. Specifically, the molecule was searched in five different subspaces, containing 1000, 5000, 10,000, 50,000 and 100,000 conformations. Since the time consumption, as defined in Section 2.3, is related to the production of knotted conformers at each step, the content in mathematical and physical knots for each of these subspaces is shown in Table 1. In order to evaluate the improvement in efficiency when GMK is deployed in a grid environment (namely the EGEE infrastructure), each experiment was split in n processes (jobs). Therefore, an experiment is defined by two parameters; the number k of conformations in the subspace and the number n of processes it is split into. This procedure determines that each process in an experiment evaluates k/n conformations. Finally, in order to obtain a good representation of the actual time consumption for the procedure, each experiment is repeated three times. Early results of this analysis were presented in [37]. The processing time in each case versus the sequential approach is presented in Figure 5.

As demonstrated in the log–log plot diagrams of Figure 5, there seems to exist a linear correlation φ between the execution time t and the number of conformations N. Based on this assumption, it is possible to write

$$\ln(t) = \varphi \times \ln(N) = \ln(N^{\varphi}) \Rightarrow t = N^{\varphi}. \qquad (8)$$

Experimentally, in the sequential approach constant, φ varies from 1.55 to 2.15. In the G-MolKnot approach, φ takes values from 1.3 to 1.5 in the worst case and from 0.35 to 0.5 in the best case. This translates into a significant improvement in the order of the time complexity. There is also a significant gap between best- and worst-case speedups. Moreover, a trend can be observed, where the speedup tends to increase with the number of jobs. Analysis of these initial experiments showed that the cause of the differences in speedup is the heterogeneity of the resources in a grid environment. Based on this conclusion, the same experiment was performed on a subset of the grid resources available, namely homogenous resources with CPU clock over 300 GHz and 4 Gb of RAM. Although there was a substantial improvement in the range of processing times (best/worst/average), especially when 20 or 40 jobs were utilised, the variations in the φ value for different number of conformations N remain. However, this can be attributed to the fact that there is a different content of knotted conformations in each subspace, therefore significantly affecting the processing time (results available on request).

Table 1. The number of produced knotted conformers after each filter for the different subspaces. Note the correlation between the output of the mathematics and the MM filters.

No of conformations	# pass math filters	# pass MM filter
1000	23	5
5000	78	7
10,000	125	11
50,000	403	26
100,000	425	29

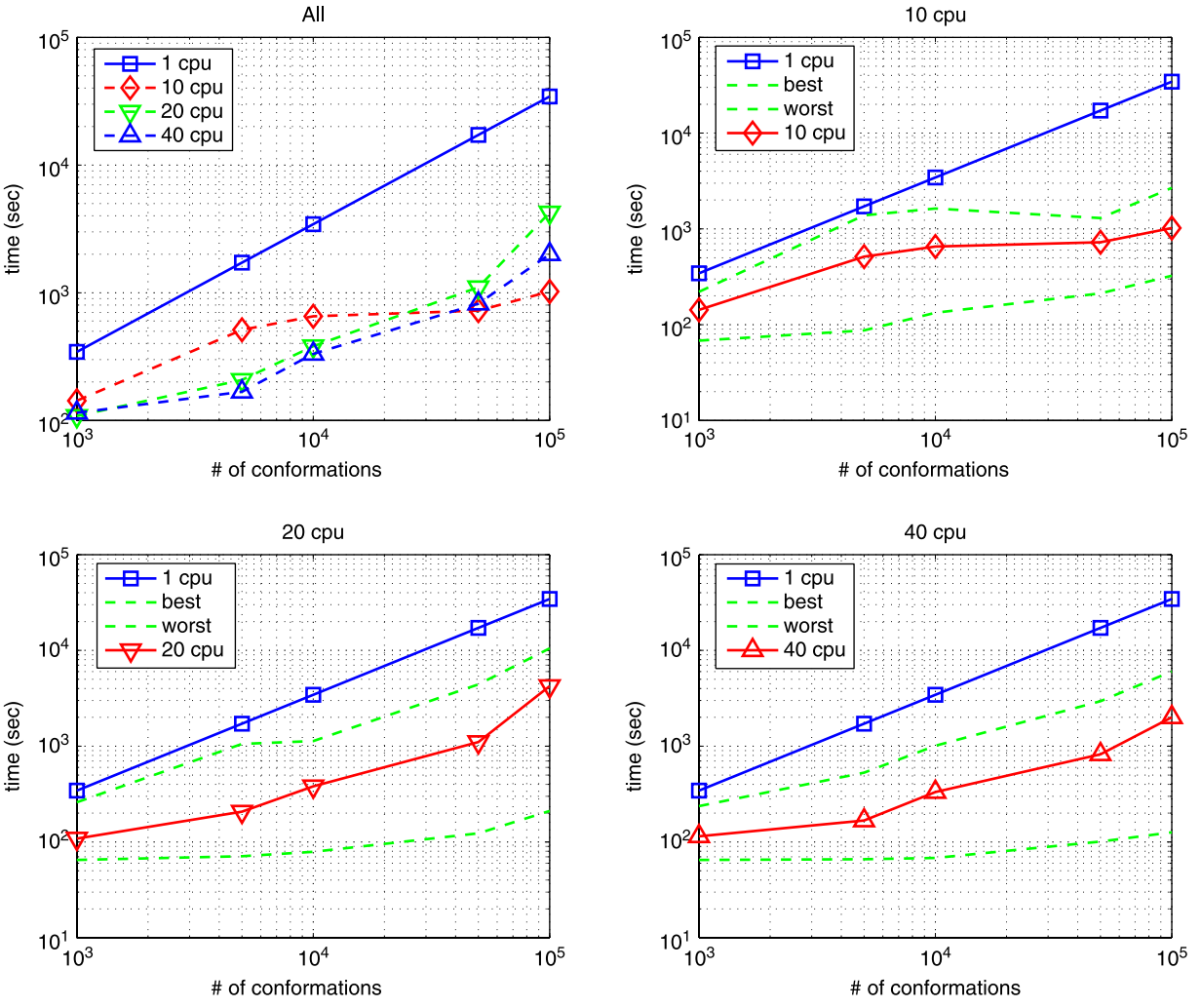

Figure 5. The first diagram shows the average performance of the algorithm in 1, 10, 20 and 40 CPUs. In all other diagrams, the red line is the average performance in each case, with the green-dotted lines above and below showing the worst and best processing time, respectively. In a homogeneous experiment, the dotted lines converge to the average line.

Regarding hardware requirements, worth noting is the following issue: the program, and more specifically the knot algorithm, was initially designed to run on 32-bit architectures. However, a non-transparent update on the grid infrastructure hardware changed several nodes to 64 bit, thus leading to erroneous results. For this reason, the program has been modified to run without problems on both architectures.

3.3 Large-scale computational experiment

3.3.1 Time efficiency

The final experiment was split in 100 different processes (jobs), each exploring a total of 5,538,440 conformations. Figures 6 and 7 show the overall statistics of the $C_{36}H_{74}$ run. Specifically, Figure 6 shows the processing time for each of the 100 jobs. Using the simplistic assumption that the sequential processing time would be the sum of the concurrent runs, the speedup measured was approximately 18 (in order to give the magnitude of the calculation, the sequential processing time would be 10.43 months, against the 16.72 days of the worst run). The calculated speedup is much greater if measured against the projected sequential time; based on the experimental results on smaller sequential runs of the same length (i.e. 36 carbon atoms in the molecular chain), the average processing time for each conformer is 0.258 s. For the TCS, this time would lead to a total processing time of 4.6 years and the corresponding speedup would be ≈ 100.

At this point, it must be noted that the processing time for each conformer through the mathematics filter is ≈ 0.03 and ≈ 10.3 s for the MM filter. Since every conformation must be examined by the math filter, this

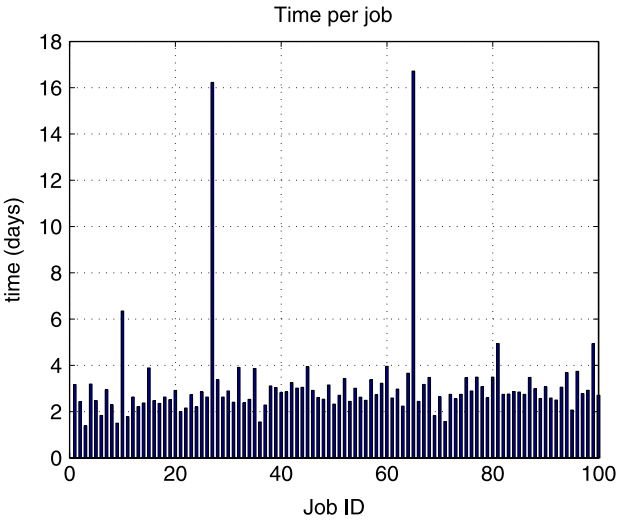

Figure 6. Time (in days) that was required for each one of the 100 jobs (i.e. splits of the conformational space). It is interesting to note that although the average processing time was 3.2 days, the distribution of the execution times is far from uniform. Specifically, the jobs with IDs 26 and 64 took a lot of time, which is explained by the greater yield of knotted conformers in these search areas (as shown in Figure 7).

step constitutes the actual bottleneck in the algorithm's performance.

Finally, the heterogeneity in the time response, as shown in Figure 6, can be explained by the nature of the actual problem, i.e. the systematic generation of conformers' produces dense and sparse areas. These dense areas are responsible for the computational peaks shown in the figure.

3.3.2 *Knotted conformers*

Based on the total output of the G-MolKnot application, the main focus shifted toward the analysis of the four-crossing knots. Regarding the three knots, there exist both loose and tight conformations. The difference between 'loose' and 'tight' lies in the deformation of the molecular chain, and therefore in the energy content of the knotted molecular conformation. The same principle applies in the knots with four crossings. However, loose four knots require larger segment lengths, and thus a longer molecular chain. For this reason, although the main production of the algorithm has been three knots, the 36 carbon atoms in the chain are not enough to create loose four knots and the production has been naturally restricted to tight four knots (as detected by the MM filter).

The main production of tight four knots ($\approx 30\%$) was delivered by the search areas of job IDs 64 and 26 (as presented in Figure 7) with 129 and 63 four knots correspondingly. The segment lengths in these areas are dominated by the combinations of 2–6–1–0 and 3–4–2–0 (overall coverage $\approx 72\%$), where the numbers in each case correspond to lengths of 1, 2, 3 and 4 (i.e. 2–6–1–0 stands for 'two segments with length 1, six segments with length 2, one segment with length 3 and none with length 4'). It is obvious that the aforementioned combinations represent a much smaller percentage of SubKCS (Equation (7)). This information can be utilised for the redesign of the algorithm in order to make it more efficient for longer molecular chains and, therefore, higher complexity knotted conformations.

At this point, there is the opportunity to discuss the problem of the production of three and four knots by the math filter and the corresponding results of the MM filter application. Regarding the math filter, subspace areas (jobs) with IDs 26 and 64 are rich in knot production (in order of $\approx 10^5$, whereas the relatively poor subspace is ID 23 (with a yield $\approx 2 \times 10^3$). Accordingly, Figure 7 shows that four knot production is at a maximum ($\approx 2 \times 10^3$) in areas 26 and 64 (as discussed previously), and at a minimum in area 23. Finally, the MM filter displays maximum production of three and four knots in the same areas as the math filter (order of $\approx 2.5 \times 10^3$ and 150, respectively).

Ordering the subspaces by production of three and four knots, as shown in Figure 8, it is obvious that approximately 20% of the jobs are enough to produce almost 50% of the knots, regarding the mathematics filter. For the MM filter, and especially for four knots, less than 10% of the jobs can produce more than 50% of the final knots. This analysis shows that there exist combinations of line segments lengths with greater propensity to creating knots. Indeed, the approach is also verified by the mathematical theorems proposed by Cantarella and Johnston [19] and Clark and Venema [20], who studied the potential of polygonal lines consisting of five and six segments to create knots. For instance, Clark and Venema [20] proved that figure-eight knot should satisfy the following inequalities (see Figure 3):

$$L_1 \geq L_2 + L_3 + L_4 + L_5.$$

$$L_6 \geq L_2 + L_3 + L_4 + L_5.$$

So instead of building all combinations, in order to reduce heterogeneity, we could enter conformers that comply with the 3.6 Theorem [20] for the case of six straight segments. Future mathematical analysis would provide us with insights for knotted conformers of more than 6 line segments.

This test shows that with the appropriate posteriori analysis of the possibilities to represent the polygonal line while keeping constant the total length and varying the

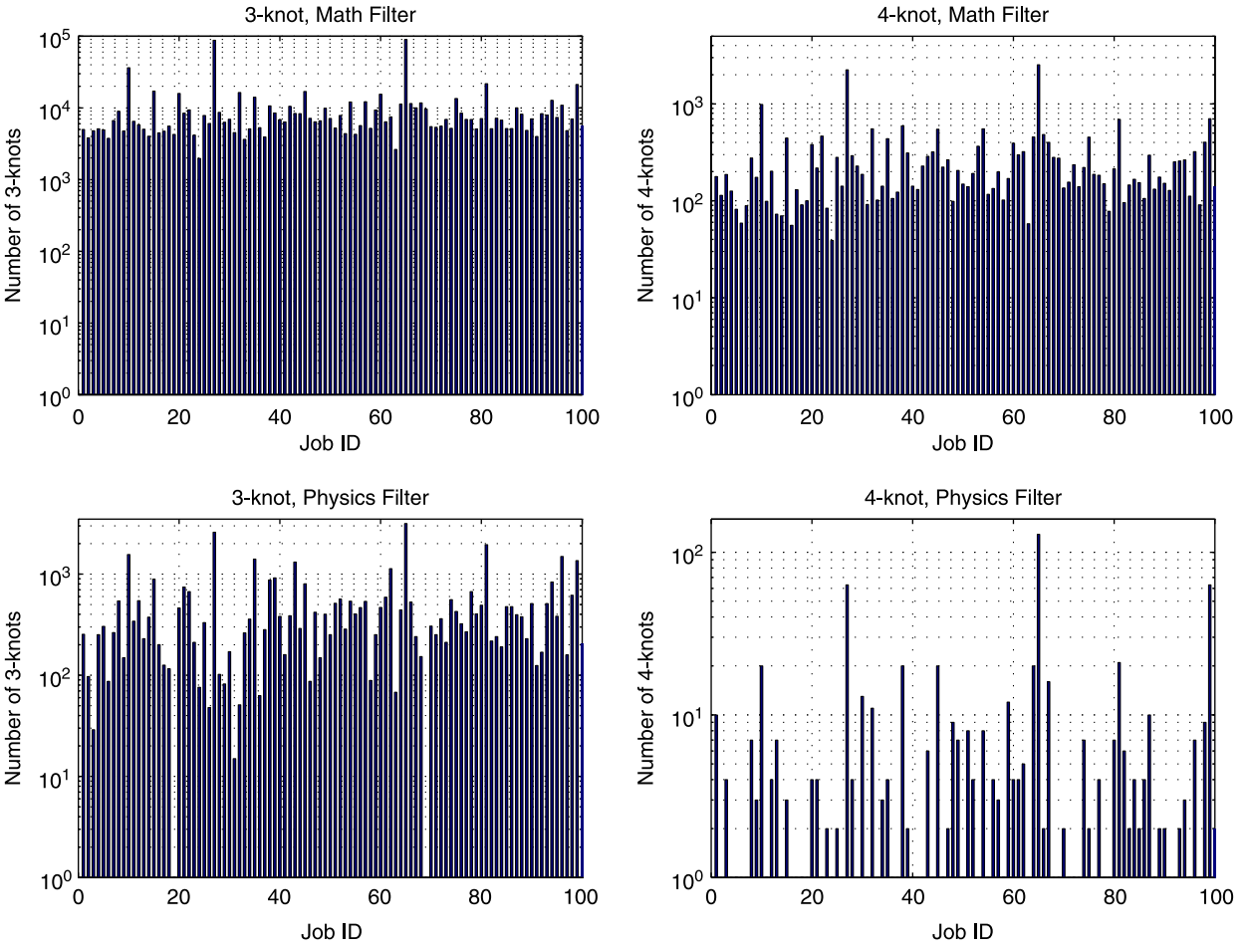

Figure 7. The top figures present (in logarithmic scale) the production of three- and four-crossing knots after the mathematics filter. The bottom figures present the production of three and four knots after the MM filter.

segments and their relative positions, it is possible to detect the combinations of the segments in order to produce 50% of the total results with only 20% of the initial computational cost. This analysis would aim to identify key patterns in the dihedral representation of the produced knotted conformers, which can be later applied in determining knottiness in molecules of greater lengths. Regarding the pattern identification process, classical data mining techniques can be employed such as association rule mining.

Figure 7 shows the number of three knots and four knots that passed the mathematics and the MM filter. Looking at the actual results of the experiment, there is some useful information regarding the yield of three knots and four knots. This information is summarised in Table 2. Krinas and Demetropoulos [27] showed that the knotted energy is spreading all over the knotted region of the molecule and induces distortion in bond lengths, bond angles and torsion angles; furthermore, a mathematical expression of the energy partition is presented. There is no relationship between the number of crossings and the knotted energy. However, it would be possible to derive one, once five knots and six knots are produced. Moreover, a quantitative structure–activity relationship could emerge by measuring the ability of tight knots to act as mediators in reactions (catalysts).

From Table 2, and given that the size of SubKCS is 553,844,736, approximately 0.2% of the conformers are three and four knots detected by the math filter. Moreover, the analysis of the results showed that there exists a possibility for further improvement, if the subspaces that are much more efficient in knot production than others can be identified. Therefore, the analysis of all subspaces, similar to the analysis of the four knots discussed previously, is expected to allow for segment length combinations that will improve the overall efficiency of the MolKnot algorithm, possibly by one or even two orders of magnitude.

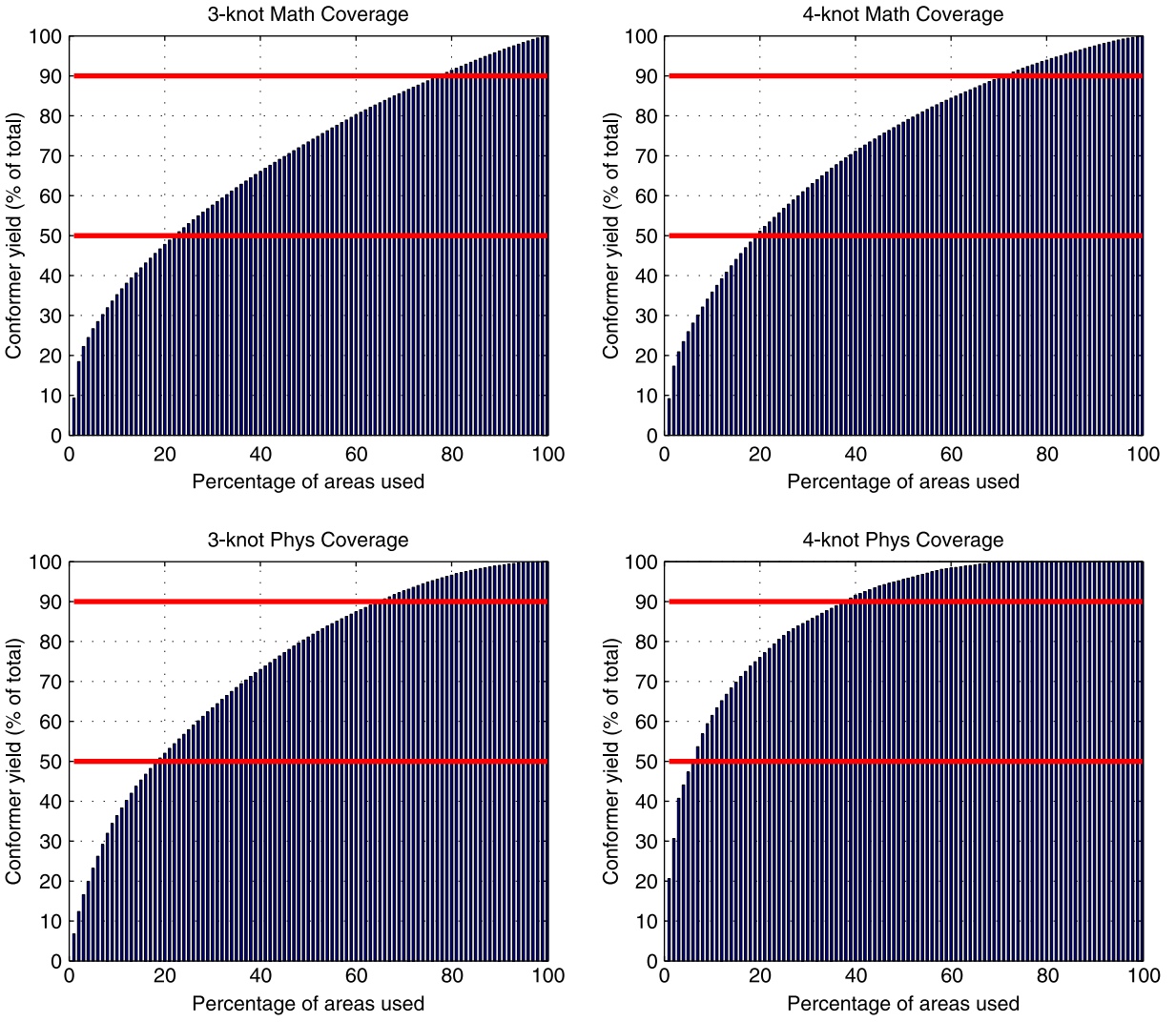

Figure 8. The diagrams present the yield in three and four knots for each of the filters. This is achieved by reordering the jobs by their yield in conformers in descending order. Thus, for example, 20% of the total jobs (computational time) produces approximately 75% of the four knots in the MM filter.

Finally, the sizable difference between the processing times of the math and the MM filters (approximately 0.03 s vs. 10.3 s) suggests that improvements can be made regarding the load balancing of the procedure. The GMK is essentially a modular algorithm as far as the two filters are concerned, hence functional decomposition can be achieved. A minor restructuring of the algorithm may allow for the step-by-step execution of the program, where the output of each step is uniformly distributed in the concurrent processes of the next one.

Table 2. Total yield of conformers after each filter, according to knot complexity (three and four crossings).

Type of knot	# pass math	# pass MM	# with energy >450[c]	# with $450 <$ energy < 1000[c]
Three knots	952,222	73,608	1,292	228
Four knots	27,326	626[a]	210[a]	64(58)[b]

[a] The numbers correspond only to the four knots produced after the classical MM filter. There are some true four knots produced after the mathematical filter, but their number is negligible compared to the MM filter number.
[b] Only 58 out of the 64 produced knots were unique.
[c] Energy units Kcal/mol.

4. Conclusions

This paper has described the development and analysis of the grid version of the MolKnot program and has shown how it has used to obtain for the first time complex structures as open molecular 4_1 knotted $C_{36}H_{74}$ conformers. During this process, it has been shown that:

- the algorithm can be transferred to the grid, due to its embarrassingly parallel nature;
- the complexity analysis of both GMK and G-MolKnot algorithms has determined that the math filter is the most CPU consuming process, and the final complexity is

$$O(nc, N, p) = \frac{\ln^{2/3}(N) \times O(nc^3)}{p}$$

- the $C_{36}H_{74}$ test case shows that the yield of the actual knot conformers is definitely not uniform across the different jobs.

There is no doubt that there exists an increasing interest for complex (with five or more crossings) molecular structures, which exhibit several interesting properties. However, the computational cost, as the actual molecular chain lengths get longer, becomes prohibitively large. An estimate of the minimum number of carbons involved in a knot formation is 30 for a four-crossing knot, 38 for five-crossing knots and 44 for the six-crossing knots. Although this hypothesis needs further testing, these estimates lead to enormous conformational spaces, over a trillion in number. Future work should target the optimisation of the G-MolKnot code in order to make it more efficient and cost effective for longer molecular chains. The next milestone should be the detection of five-crossing knots.

An additional issue that requires further study is the heterogeneity in time efficiency, due to the nature of the problem as discussed in Section 3.3.1. Adaptive loading techniques may provide satisfactory solutions to this problem.

In the end, the adaptation of the MolKnot code to a grid environment has enabled the detection of interesting structures with four crossings, and has provided an analysis framework for more efficient searching of knots in longer molecular chains.

Acknowledgements

The research Project is co-funded by the European Union – European Social Fund (ESF) & National Sources, in the framework of the program 'HRAKLEITOS' of the 'Operational Program for Education and Initial Vocational Training' of the 3rd Community Support Framework of the Hellenic Ministry of Education. The authors gratefully acknowledge the computer time provided by the Hellas Grid Infrastructure and the computer time provided by the Research Center for Scientific Simulations (RCSS) of Ioannina University, Greece. The following software packages have been used: Knot [38]; Tinker [39]; and Gamess-US [40].

Notes

1. Email: fpsom@issel.ee.auth.gr
2. Email: mitkas@eng.auth.gr; http://issel.ee.auth.gr/en/mitkas
3. Email: me00599@cc.uoi.gr

References

[1] O. Lukin and F. Vögtle, *Knotting and threading of molecules: chemistry and chirality of molecular knots and their assemblies*, Angew. Chem. Int. Ed. Engl. 44 (2005), pp. 1456–1477.
[2] C. Liang and K. Mislow, *Knots in proteins*, J. Am. Chem. Soc. 116 (1994), pp. 11189–11190.
[3] W.R. Taylor, *A deeply knotted protein structure and how it might hold*, Nature 406 (2000), pp. 916–919.
[4] L.F. Liu, R.E. Depew, and J.C. Wang, *Knotted single-stranded DNA rings: a novel topological isomer of circular single-stranded DNA formed by treatment with* Escherichia coli *ω protein*, J. Mol. Biol. 106 (1976), pp. 439–452.
[5] H. Adams, E. Ashworth, G.A. Breault, J. Guo, C.A. Hunter, and P.C. Mayers, *Knot tied around an octahedral metal centre*, Nature 411 (2001), p. 763.
[6] V. Balzani, A. Credi, F.M. Raymo, and J.F. Stoddart, *Artificial molecular machines*, Angew. Chem. Int. Ed. 39 (2000), pp. 3348–3391.
[7] O. Lukin, T. Kubota, Y. Okamoto, F. Schelhase, A. Yoneva, W.M. Müller, U. Müller, and F. Vögtle, *Knotaxanes–rotaxanes with knots as stoppers*, Angew. Chem. Int. Ed. 42 (2003), pp. 4542–4545.
[8] F.M. Raymo and J.F. Stoddart, *Organic Template-directed syntheses of catenanes, rotaxanes, and knots*, in *Molecular Catenanes, Rotaxanes and Knots*, J.-P. Sauvage and C. Dietrich-Buchecker, eds., Wiley-VCH, Weinheim, 1999, pp. 143–176.
[9] G. Bottari, F. Dehez, D.A. Leigh, P.J. Nash, E.M. Pérez, J.K.Y. Wong, and F. Zerbetto, *Entropy-driven translational isomerism: a tristable molecular shuttle*, Angew. Chem. Int. Ed. 42 (2003), pp. 5886–5889.
[10] D.J. Craik, M. Čemažar, and N.L. Daly, *The cyclotides and related macrocyclic peptides as scaffolds in drug design*, Curr. Opin. Drug Discov. Devel. 9 (2006), pp. 251–260.
[11] E.K. Wilson, *Chemists turned visionaries*, Chem. Eng. News 78 (2000), pp. 39–45.
[12] A.V. Vologodskiĭ, A.V. Lukashin, M.D. Frank-Kamenetskiĭ, and V.V. Anshelevich, *The knot problem in statistical mechanics of polymer chains*, Pisma v Zhurnal Eksperimentalnoi i Teoreticheskoi Fiziki 66 (1974), pp. 2153–2163 [Sov. Phys. J. Exp. Theor. Phys. Lett. (JETP Lett.), 39 (1974), pp. 1059–1063].
[13] K. Koniaris and M. Muthukumar, *Knottedness in ring polymers*, Phys. Rev. Lett. 66 (1991), pp. 2211–2214.
[14] A.M. Saitta, P.D. Soper, E. Wasserman, and M.L. Klein, *Influence of a knot on the strength of a polymer strand*, Nature 399 (1999), pp. 46–48.
[15] P. Virnau, Y. Kantor, and M. Kardar, *Knots in globule and coil phases of a model polyethylene*, J. Am. Chem. Soc. 127 (2005), pp. 15102–15106.
[16] K. Millet, A. Dobay, and A. Stasiak, *Linear random knots and their scaling behavior*, Macromolecules 38 (2005), pp. 601–606.
[17] C.S. Krinas and I.N. Demetropoulos, *A systematic algorithm capable to yield open molecular knots: application to alkanes, polyethylene oxides and peptides*, Chem. Phys. Lett. 433 (2007), pp. 422–426.
[18] Generalized Molecular Knotting Algorithm, http://users.uoi.gr/me00599/index.html
[19] J. Cantarella and H. Johnston, *Nontrivial embeddings of polygonal intervals and unknots in 3-space*, J. Knot Theory Ramificat. 7 (1998), pp. 1027–1039.

[20] T.J. Clark and G.A. Venema, *Classifying polygonal chains of six segments*, J. Knot Theory Ramificat. 13 (2004), pp. 479–514.
[21] D. Bonchev and D.H. Rouvray, *Chemical Graph Theory: Introduction and Fundamentals*, Gordon and Breach Science Publishers, New York, 1990.
[22] M. Tasioula-Margari and I.N. Demetropoulos, *Viscosity–structure relationship of dilute triglycerides' solutions correlation to retention time in reversed-phase liquid chromatography*, J. Am. Oil Chem. Soc. 69 (1992), pp. 1112–1117.
[23] A.C. Wallace, R.A. Laskowski, and J.M. Thornton, *Derivation of 3D coordinate templates for searching structural databases: application to the SerZHisZAsp catalytic triads of the serine proteinases and lipases*, Protein Sci. 5 (1996), pp. 1001–1013.
[24] V.A. Tatsis, A. Stavrakoudis, and I.N. Demetropoulos, *Lysine-based TrypsinActSite(LysTAS): a configurational tool of the TINKER software to evaluate lysine based branched cyclic peptides as potential chymotrypsin-mimetics*, Mol. Simul. 32 (2006), pp. 643–644.
[25] F.M. Menger, *An alternative view of the enzyme catalysis*, Pure Appl. Chem. 77 (2005), pp. 1873–1886.
[26] J.W. Alexander, *Topological invariants of knots and links*, Trans. Amer. Math. Soc. 30 (1928), pp. 275–306.
[27] C.S. Krinas and I.N. Demetropoulos, *$C_{22}H_{46}$: the smallest open 3_1 knotted alkane by computer-aided design*, J. Mol. Graph. Model. 26(7) (2008), pp. 1153–1159.
[28] L. Philpott, *Pocket guide to Knots*, New Holland Publishers, London, 2006.
[29] M. Delbrück, *Knotting problems in biology*, Proc. Symp. Appl. Math. 14 (1962), pp. 55–63.
[30] I. Foster and C. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publishers, San Fransisco, 1998.
[31] Enabling Grids for E-sciencE, http://www.eu-egee.org
[32] Uniform Interface to Computing Resources, http://www.unicore.eu/
[33] The Globus Alliance, http://www.globus.org/
[34] Research Center For Grid and Service Computing, http://vega.ict.ac.cn/
[35] H.E. Polychroniadou, F.E. Psomopoulos, and P.A. Mitkas, *G-Class: a divide and conquer application for grid protein classification*, in *Proceedings of the 2nd ADMKD 2006: Workshop on Data Mining and Knowledge Discovery (in conjunction with ADBIS 2006: The 10th East-European Conference on Advances in Databases and Information Systems*, 2006, pp. 121–132.
[36] D. Coppersmith and S. Winograd, *Matrix multiplication via arithmetic progressions*, J. Symb. Comput. 9 (1990), pp. 251–280.
[37] F.E. Psomopoulos, P.A. Mitkas, C.S. Krinas, and I.N. Demetropoulos, *G-MolKnot: a grid enabled systematic algorithm to produce open molecular knots*, presented at the 1st HellasGrid User Forum, 2008.
[38] B.A. Harris and S.C. Harvey, *Program for analyzing knots represented by polygonal paths*, J. Comput. Chem. 20 (1999), pp. 813–818.
[39] J.W. Ponder and F.M. Richards, *An efficient Newton-like method for molecular mechanics energy minimization of large molecules*, J. Comput. Chem. 8 (1987), pp. 1016–1024.
[40] M.W. Schmidt, K.K. Baldridge, J.A. Boatz, S.T. Elber, M.S. Gordon, J.H. Jensen, S. Koseki, N. Matsunaga, K.A. Nguygen, S. Su, T.L. Windus, M. Dupuis, and J.A. Montgomery, *Generic atomic and molecular electronic structure system*, J. Comput. Chem. 14 (1993), pp. 1347–1363.

Appendix A

Figure 3(b). Molecular segments used for the assembly of the initial molecular geometry (L_i, black colour), (B_j, red colour) with: $L_1 = \Phi_1$, $B_1 = \Phi_2\Phi_3$, $L_2 = \Phi_4$, $B_2 = \Phi_5\Phi_6$, $L_3 = \Phi_7\Phi_8$, $B_3 = \Phi_9\Phi_{10}$, $L_4 = \Phi_{11}\Phi_{12}$, $B_4 = \Phi_{13}\Phi_{14}$, $L_5 = \Phi_{15}\Phi_{16}$, $B_5 = \Phi_{17}\Phi_{18}$, $L_6 = \Phi_{19}\Phi_{20}\Phi_{21}$, $B_6 = \Phi_{22}\Phi_{23}$, $L_7 = \Phi_{24}$, $B_7 = \Phi_{25}\Phi_{26}$, $L_8 = \Phi_{27}\Phi_{28}$, $B_8 = \Phi_{29}\Phi_{30}$ and $L_9 = \Phi_{31}\Phi_{32}\Phi_{33}$, where $\Phi_1, \Phi_2, \ldots, \Phi_{33}$ are CCCC torsion angles in the carbon backbone of the $C_{36}H_{74}$ molecule.

Figure 3(c). The initial building sequence $\Phi_1, \Phi_2, \ldots, \Phi_{33}$ is: 180°, +95°, −60°, 180°, −95°, +60°, 180°, 180°, +95°, −60°, 180°, 180°, +95°, −60°, 180°, 180°, −60°, +95°, 180°, 180°, 180°, +60°, −95°, 180°, −95°, +60°, 180°, 180°, −95°, +60°, 180°, 180° and 180°.

Figure 3(d). Tube representation of the outcome of the geometry optimisation procedure of the $C_{36}H_{74}$ hydrocarbon; the sequence of CCCC backbone torsion angles (at B3LYP level): −174.6°, 133.1°, 38.1°, −5.9°, 58.7°, −99.0°, 132.5°, −133.9°, 176.2°, −130.5°, 142.6°, −137.1°, 127.7°, −53.1°, −81.1°, 91.9°, −65.6°, 84.8°, −122.4.0°, −103.6°, 50.2°, 58.7°, −92.9°, 111.7°, −95.1°, 115.8°, −127.8°, 129.0°, −163.9°, 109.6°, −172.9°, 56.4° and −161.6° (hydrogens are not shown for clarity). CC bonds (BL) are in the range 1.538 Å ≤ BL ≤ 1.685 Å and CCC bond angles (BA) are between 109.3° ≤ BA ≤ 145.0° (B3LYP/6-31G** optimised geometry. Frequency calculations confirmed the local minimum at the same level of theory).