A Novel Method for Multiple Sequence Alignment Based on Wavelet Package Transform

Junfeng Gu[1] , Xicheng Wang[1] , Jincheng Zhao[2]

[1] The State Key Laboratory of Structural Analysis for Industrial Equipment, Dalian University of Technology, Dalian 116024, China
[2] Institute of Bioinformatics and Molecular Design, Dalian University, Dalian 116622, China

valley@student.dlut.edu.cn

Abstract

Multiple sequence alignment is one of the essential tools of studying bioinformatics. It plays an important role in the evolution analysis and protein structure prediction. A novel method for multiple sequence alignment based on wavelet package transform (MAWPT) is developed in this paper. By means of wavelet package analysis, homologous regions can be found rapidly. Numerical examples show that it has good alignment accuracy and gives high efficiency.

Keyword: Bioinformatics, multiple sequence alignment, wavelet package transform, BAliBASE

I. Introduction

A central theme of modern molecular biology is to elucidate the interrelationships among genetic information, higher-order structures of gene products, and their biological functions. Multiple sequence alignment (MSA) is one of the essential tools of studying bioinformatics. It plays an important role in the evolution and function analysis of protein family, as well as in structure prediction of protein and RNA. By means of MSA, the similarity relationships among the several sequences can be obtained, and then we can find out the basic features of a protein family and identify motifs preserved by evolution.

MSA can be stated as an optimization problem, and the optimal alignment can be obtained by optimizing a scoring function. MSA can be classified in three categories: exact, progressive and iterative. Exact algorithms are high quality heuristics that deliver an alignment usually very close to optimality, sometimes but not always within well-defined boundaries. They are suitable for no more than 30 sequences only. Progressive alignments are by far the most widely used. They depend on a progressive assembly of the multiple alignment where sequences or alignments are added one by one so that never more than two sequences (or multiple alignments) are simultaneously aligned using dynamic programming. This approach has the great advantage of speed and simplicity combined with reasonable sensitivity, even if it is by nature a heuristic that does not guarantee any level of optimization. CLUSTAL W, developed by Thompson [1], is a standard progressive method for multiple sequence alignment. Iterative alignment methods depend on algorithms able to produce an alignment and to refine it through a series of cycles until no more improvements can be made. Iterative methods can be deterministic or stochastic, depending on the strategy used to improve the alignment. The simplest iterative strategies are deterministic, such as PRRP/PRRN developed by GOTOH [2]. They involve extracting sequences one by one from a multiple alignment and

realigning them to the remaining sequences. Stochastic iterative methods include HMM training [3] and simulated annealing [4,5] or genetic algorithms [6,7,8,9]. The main advantage is to allow for a good conceptual separation between optimization processes and objective function. Recently, some algorithms that combine several strategies have been developed to improve the alignment efficiency, for example, MAFFT combines Fourier transform and iterative strategy [10], and MUSCLE combines iterative and progressive strategy [11], etc.

To improve both the accuracy and speed is a difficult problem. The adoption of anchor point is a better way to resolve this problem. Anchor point is the high similarity segments among several sequences, which corresponds to the preserved regions of a protein structure, e.g., secondary structure or hydrophobic core. Accurate locating of these regions will not only improve the accuracy of MSA, but also decrease the alignment time and contribute to help analyzing the relations between protein sequence and structure. MAFFT mentioned above depends on Fourier transform to find out high similarity regions, and has got favorable results.

Wavelet is an efficiency tool developed for digital processing in recent twenty years, and wavelet has been used in signal analysis, image processing and nonlinear studying, etc [12]. Corresponding to Fourier analysis, wavelet analysis is efficient for multi-resolution analysis and local feature analysis of a signal, so it is very convenience to study non-stationary signal. It involves decomposing a given signal into its scale and space components. Information can be obtained about both the amplitude of any periodic signal as well as when/where it occurred in time/space. Wavelet analysis thus localizes both in space and scale, unlike the Fourier transform in which time/scale information is lost. Recently, the use of both types of wavelet transform, continuous (CWT) and discrete (DWT) in the bioinformatics field is promising. Continuous wavelet transform allows a one-dimensional signal to be viewed in a more discriminative two-dimensional time-scale representation (scalogram). Rough classification of proteins mediated by the study of three levels of such scalograms of hydropathy data has been proposed [13]. Detection of repeats of particular secondary or supersecondary structural units is another application of continuous wavelet transform [14]. Discrete wavelet transform has been applied to hydrophobicity signals in order to predict hydrophobic cores in proteins [15]. Protein sequence similarity has also been studied using DWT of a signal associated with the average energy states of all valence electrons of each amino acid [16]. Wavelet package is the extension and development of wavelet, and it gives a valuable way to more suitable cross-cutting of time-frequency space. In this article, a novel multiple sequence alignment method based on wavelet package transform has been developed, it can detect high similarity regions quickly. The speed and accuracy are tested by simulation program ROSE and BAliBASE benchmark, Compared with several existing methods, the results show the effectiveness of this method.

Table 1. Polarity value and volume value of amino acids【Grantham，1974】

	Polarity	**Volume**		**Polarity**	**Volume**
Ala	8.1	31.0	**Lue**	4.9	111.0
Arg	10.5	124.0	**Lys**	11.3	119.0
Asn	11.6	56.0	**Met**	5.7	105.0
Asp	13.0	54.0	**Phe**	5.2	132.0
Cys	5.5	55.0	**Pro**	8.0	32.5
Gln	10.5	85.0	**Ser**	9.2	32.0
Glu	12.3	83.0	**Thr**	8.6	61.0
Gly	9.0	3.0	**Trp**	5.4	170.0
His	10.4	96.0	**Val**	6.2	136.0

Ile	5.2	111.0	**Tyr**	5.9	84.0

II. MSA method based on WPT

2.1 Transfer a protein sequence to digital signal sequence

Protein sequence is composed of 20 kinds of amino acids, each of which has a distinct physico-chemical property. In the evolution history of proteins, substitutions between physico-chemically similar amino acids tend to preserve the structure of proteins, and such neutral substitutions have been accumulated in molecules during evolution. It is therefore reasonable to consider that an amino acid can be assigned to a vector whose components are the volume value and the polarity value introduced by Grantham [17], thus a protein sequence is transferred to a digital sequence. Table 1 shows the volume value and the polarity value of 20 kinds of amino acids.

2.2 Wavelet package transform

Wavelet package was developed by M.V.Wickerhauser and R.R.Coifman based on wavelet analysis, and can be realized as gradually cross-cutting of function space. When we do wavelet decomposition to a signal, the high frequency signal and low frequency signal will take up half width of the signal frequency space respectively. When decomposition was down again, the low frequency part can be divided into two same width frequency spaces. During the process of wavelet package decomposition, not only the low frequency part is gradually decomposed, but also the high frequency is decomposed further more, so we can obtain more accurate local information.

In wavelet package decomposition, we can unify scale subspace and wavelet subspace by a new subspace, which can be denoted as

$$U_j^n = close\{2^{-j/2} w_n(2^{-j}t-k), k \in Z\},\ j \in Z, n \in Z_+ \qquad \textbf{(1)}$$

it is a subspace of $L^2(R)$ space composed of standard Orthogonal basis $\{2^{-j/2} w_n(2^{-j}t-k), k \in Z\}$.

The wavelet package transform coefficients of signal $x(t)$ are denoted as

$$\left\{C_p^{j.n} \mid p \in Z\right\}, \left\{C_p^{j,n} = \int_{-\infty}^{\infty} x(t) 2^{-j/2} w_n(2^{-j}t-p)dt\right\} \qquad \textbf{(2)}$$

So the coefficients $\left\{C_p^{j,2n} \mid p \in Z\right\}$ and $\left\{C_p^{j,2n+1} \mid p \in Z\right\}$ of signal $x(t)$ satisfy the following equations:

$$C_p^{j,2n} = \sum_l h_{l-2p} C_l^{j-1,n} \qquad \textbf{(3)}$$

$$C_p^{j,2n+1} = \sum_l g_{l-2p} C_l^{j-1,n} \qquad \textbf{(4)}$$

Equation（3）and（4）are called wavelet package algorithm, and wavelet package decomposition process is to gradually decompose the target signal into more and more narrow frequency spaces with a group of lowpass and highpass filters that are conjugated and normalized.

Wavelet package transform converts signal from time space to time-frequency space, and from one dimension to two dimensions, so we can find out the difference between two signals more clearly. The signals that have high similarity in time space maybe is very different in time-frequency space, and this is the reason why we use two-dimensional correlation in time-frequency space for excluding noise [18]. If we use $f_1(t)$ and $f_2(t)$ to represent sequence a and b respectively, for finding out the similarity between $f_1(t)$ and $f_2(t)$, we calculate the two dimension correlation coefficient between the wavelet package transform of $f_1(t)$ with a lag i and that of $f_2(t)$. The coefficient can be calculated by the following equation:

$$v(i) = \frac{\sum_a \sum_\tau W_{f_1(t)}(a, \tau - i) W_{f_2(t)}(a, \tau)}{\sqrt{\sum_a \sum_\tau (W_{f_1(t)}(a, \tau - i))^2, \sum_a \sum_\tau (W_{f_2(t)}(a, \tau))^2}} \quad \textbf{(5)}$$

When $f_1(t)$ with a lag i and $f_2(t)$ have a high similarity region in time-frequency space, we can get a peak of $v(i)$. Otherwise, even if there is heavy noise existing, because of the high difference between two signals, we can get a low value of $v(i)$. Polarity value sequence and volume value sequence are dealt with respectively, and are added up. Finally, we obtain a series correlation value as varying with the value of lag i.

If two sequences have homologous regions, the correlation coefficients of them will reach a peak through wavelet package transform, thus the lag value i can be obtained. For finding out these positions, we slide along the sequence with a window and calculate the similarity value of region in the window. If the similarity value exceeds a threshold, a homologous region can be found out, i.e. an anchor point. In this article, we deal with the top 20 peaks and set the width of sliding window 30.

2.3 To obtain an alignment by arranging homologous segments

For obtaining an alignment of two sequences, the detected segments must be arranged reasonably. Dynamic programming algorithm is applied to arrange these segments. At first, we need to construct a similarity matrix S_{ij} （$1 \le i, j \le n$， n is the number of similarity segments）. If the i th segment of sequence a and j th segment of sequence b are homologous segments, then the value of S_{ij} is 1, otherwise, it's zero. After we get the most reasonable arrangement through dynamic programming, the global similarity matrix is divided into several sub-matrices by anchor point, so we can deal with them respectively. The more similarity segments, the less time the alignment will consume.

2.4 Extension to group-to-group alignment and DNA alignment

The procedure described above can be easily extended to group-to-group alignment. $f_1(t)$ can be replaced as

$$f_{gruop1(t)} = \sum_{i \in group1} w_i \cdot f_i(t) \quad \textbf{(6)}$$

Where w_i is the weight of sequence i, which can be calculated by means of the progressive methods, such as CLUSTAL W [1].

This method can be applied to convert a DNA into a sequence of four dimension vectors whose components are the frequencies of A, T, G and C at each column, instead of volume and polarity values. In this case, correlation between two nucleotide sequences is

$$c(k) = c_A(k) + c_T(k) + c_G(k) + c_C(k) \quad \textbf{(7)}$$

2.5 Similarity matrix and gap penalty

We adopt here the normalized similarity and gap penalty developed by Katoh [12]. Similarity matrix $\hat{M}_{ab}$ can be represented as

$$\hat{M}_{ab} = [(M_{ab} - average2)/(average1 - average2)] + S^a \quad \textbf{(8)}$$

Where $average1 = \sum_a f_a M_{aa}$, $average2 = \sum_{a,b} f_a f_b M_{ab}$, M_{ab} is raw PAM similarity matrix, f_a is the frequency of occurrence of amino acids a , and S^a is a parameter that functions as a gap extension penalty. Gap penalty can be represented as

$$G(i,x) = S^{op} \cdot \{1 - [g^{start}(x) + g^{end}(i)]/2\} \quad \textbf{(9)}$$

Where S^{op} corresponds to a gap opening penalty, $g_1^{start}(x)$ is the number of the gaps that start at the x th site, and $g_1^{end}(i)$ is the number of the gaps than end at the i th site.

For the further improvement of alignment accuracy, the alignment is divided into two groups and realigned. We employ a technique called tree-dependent restricted partitioning [19]. The iterative process is repeated until no better scoring alignment is obtained.

III. Results

In order to evaluate the performance of the MAWPT method, we give comparisons with other methods by running standard alignment program CLUSTAL W and MAFFT. For evaluating the efficiency of MAWPT, simulation program ROSE is applied here. ROSE is a simulation program that can generate protein families with specified number、length and evolution distance [20]. Five protein families with sequences of length 100、200、500、800 and 1000 are generated by running ROSE, respectively, and each of them is composed of eight sequences. Figure 1 shows the efficiency comparisons with the two methods mentioned above. The MAWPT method can give faster speed than MAFFT, and the consumed time increases more slowly when the length of the protein sequences increases. CLUSTAL W has the fastest speed, but its accuracy is not good.

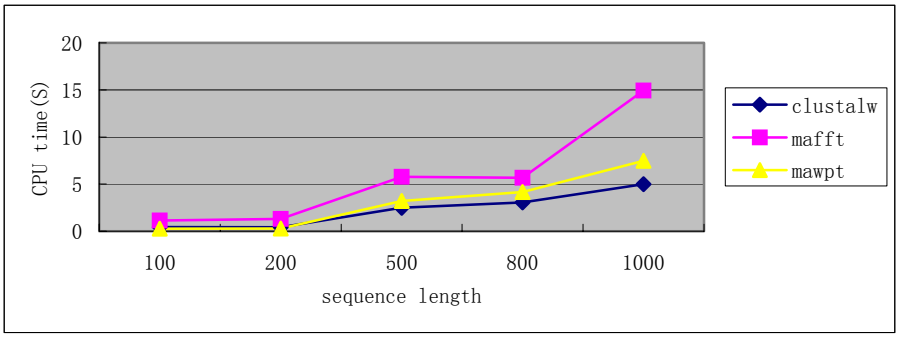
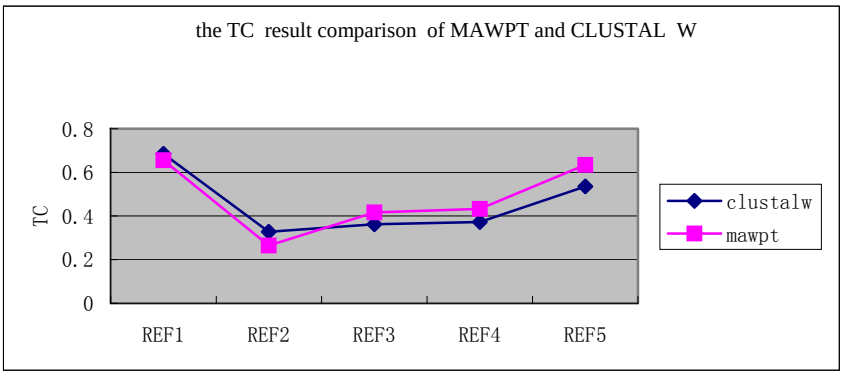


Fig. 1. Resulting profiles of CPU time compared with other two methods

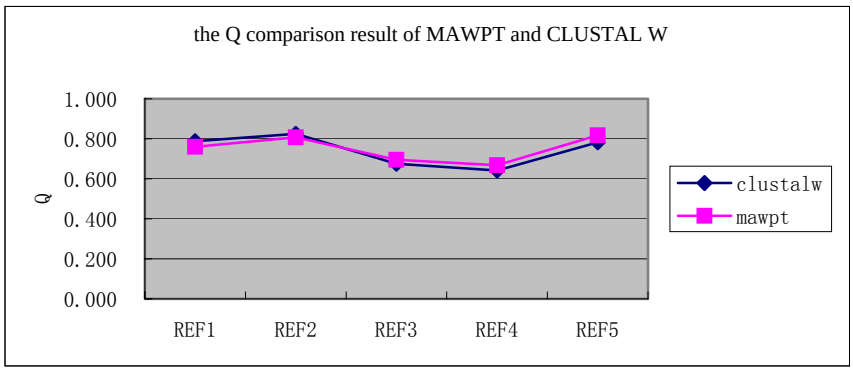
For evaluating the accuracy of MSAs, BAliBASE benchmark is here introduced, which is a database of correct alignment based on three-dimensional structural superimpositions [21]. The BAliBASE database is categorized into five different cases. The first is made up of phylogenetically equidistant members of similar length. In the second case, each alignment contains up to three orphan sequences with a group of close relatives. The third contains up to four distantly related groups, while the fourth and fifth involve long terminal and internal insertions, respectively. All these cases will be referred to as REF 1-5 hereafter.

Figure 2 and figure 3 show the comparison results with CLUSTAL W and MAFFT, respectively. In these figures, TC is the ratio of correctly aligned columns, and Q is the sum-of-pairs scores. From figure 2, compared with standard progressive alignment method CLUSTAL W, we can easily see that at the first references which is a sum of phylogenetically equidistant members of similar length, the present method get almost equal accuracy, but at REF2, it gets less accuracy, which proves that it is not so sensitive to distantly related sequences. At REF 3-5 that contain much local mutation information, the present method shows a certain extent superiority, which proves the efficiency of wavelet tool for the local mutation problem.

In figure 3, the present method gets comparable results with MAFFT, but they have some differences in the details. The results of them at REF 1 and REF 2 are consistent approximately. The

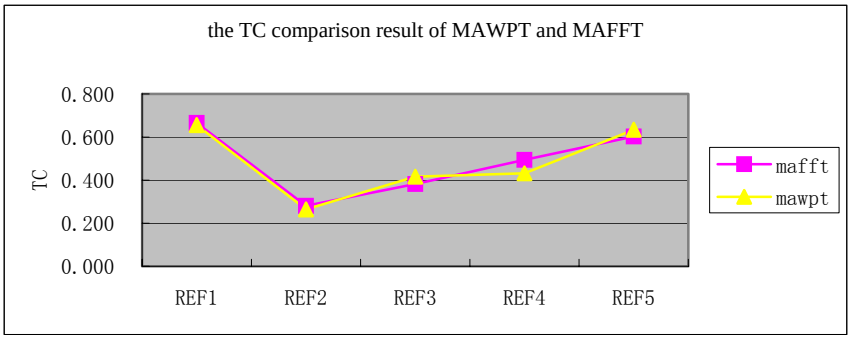


a

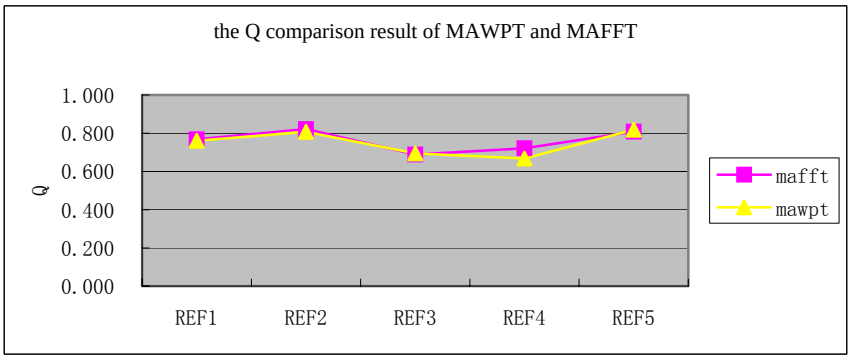


b

Fig. 2. The TC and Q comparison results of MAWPT and CLUSTAL W



a



b

Fig. 3. The TC and Q comparison results of the present method and MAFFT

sequences of REF4 which have long N/C terminal extension makes the global appearance of the sequences very different, and the present get a lower accuracy value compared with MAFFT. But at REF3 and REF5, it get a certain extent superiority, which proves the wavelet package is more efficiency in excluding noise and dealing with local mutation than Fourier transform.

IV. Conclusions

In this paper a wavelet package transform algorithm has been presented for multiple sequence alignment. It has been shown to be very rapid to find homologous regions. Numerical results have demonstrated the effectiveness of the method and compared favorably with other methods. The method is suitable for parallel computing, further work is doing for developing this method.

ACKNOWLEGDEMENTS

The authors are grateful for financial support from the Major State Basic Research Development Program of China (No. 2004CB518901) and National Natural Science Foundation of China (No.10572033).

References

[1] J. D. Thompson, et al, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic. Acids. Research*, 1994, 22, pp. 4673-4680.

[2] Osamu Gotoh, Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol*, 1996, 264, pp. 823-838.

[3] Krogh A, et al, Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol*, 1994, 235, pp. 1501-1531.

[4] J. Kim, S. Pramanik and M. J. Chung, Multiple sequence alignment using simulated annealing. *Comp. Applic. Biosci*, 1994, 10, pp. 419-426.

[5] J. Kim, J. R. Cole and S. Pramanik, Alignment of possible secondary structures in multiple RNA sequences using simulated annealing. *Comp. Applic. Biosci*, 1996, 12, pp. 259-267.

[6] L. A. Anabarasu, Multiple sequence alignment using parallel genetic algorithm, *The Second Asia-Pacific Conference on Simulated Annealing*, Canberra, Australia 1998.

[7] R. Gonzalez, Multiple protein sequence comparison by genetic algorithms, *SPIE-98*, 1999.

[8] C. Zhang and A. K. Wong, A genetic algorithm for multiple molecular sequence alignment. *Comput. Appl. Biosci*, 1997, 13, pp. 565-581.

[9] C. Notredame and D. G. Higgins, SAGA: sequence alignment by genetic algorithm. *Nucleic. Acids. Research*, 1996, 24, pp. 1515-1524.

[10] Kazutaka Katoh, et al, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic. Acids. Research*, 2002, 30, pp. 3059-3066.

[11] R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic. Acids. Research*, 2004, 32, pp.1792-1797.

[12] Yuhua Peng, Wavelet transform and engineer application (in Chinese), Science press.

[13] A. J. Mandell, K. A. Selz and M. F. Shlesinger, Wavelet transformation of protein hydrophobicity sequences suggests their memberships in structural families. *Physica. A*, 1997, 244, pp. 254-262.

[14] K. B. Murray, D. Gorse and J. M. Thornton, Wavelet transforms for the characterization and detection of repeating motifs. *J. Mol. Biol*, 2002, 316, pp.341-363.

[15] H. Hirakawa, S. Muta and S. Kuhara, The hydrophobic cores of proteins predicted by wavelet analysis. *Bioinformatics*, 1999, 15, pp. 141-148.

[16] C. H. De Trad, Q. Fang and I. Cosic, Protein sequence comparison based on the wavelet transform approach. *Protein Eng*, 2002, 15, pp. 193-203.

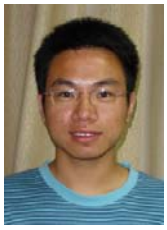
[17] R. Grantham, Amino acid difference formula to help explain protein evolution. *Science*, 1974, 185, pp. 862-864.

[18] Zhongjin Jiang, et al, Application of wavelet packet in time-delay estimation of signal of vibroseis exploration (in Chinese). *Journal of Jilin University*, 2003, 21, pp. 105-109.

[19] M. Hirosawa, et al, Comprehensive study on iterative algorithms of multiple sequence alignment. *Comput. Appl. Biosci*, 1995, 11, pp. 13–18.

[20] Jene Stoye, et al, Rose: generating sequence families. *Bioinformatics*, 1998, 14, pp.157-163.

[21] Anne Bahr, J. D. Thompson, et al, BaliBASE (Benchmark Alignment database): enhancements for repeats, transmembrance sequence and circular permutations. *Nucleic. Acids. Research*, 2001, 29, pp. 323-326.



Junfeng Gu is a student studying for his PhD in the Dalian university of technology, China.

His areas of interest include bioinformatics, computational biology, the structure and folding of protein.

Xicheng Wang received his PhD from the Dalian University of Technology, China in 1989. He is a full professor of mechanics and computer science in the Dalian University of Technology.

Professor Wang's areas of interest include applied mechanics, computer science, computational biology, optimization methods and drug molecular design. He has published more than 160 technical papers in mechanics, applied mathematics, drug molecular design and computer science.

Jincheng Zhao is a full professor of mechanics in the Dalian University, China. Now he works in the Institute of Bioinformatics and Molecular Design of Dalian University, China.

Professor Zhao's current working areas include engineering mechanics, bioinformatics and molecular design. He has published more than 50 technical papers in engineering mechanics, bioinformatics and drug molecular design.