Journal of Bioinformatics and Computational Biology


# HOMOLOGOUS SYNTENY BLOCK DETECTION BASED ON SUFFIX TREE ALGORITHMS

YU-LUN CHEN, CHIEN-MING CHEN, TUN-WEN PAI[*]

*Department of Computer Science and Engineering & Center of Excellence for the Oceans, National Taiwan Ocean University, No.2, Peining Road, Keelung, Taiwan 20224, Republic of China*
[*]*twp@mail.ntou.edu.tw*

HON-WAI LEONG[†], KET-FAH CHONG

*Department of Computer Science, National University of Singapore, 13 Computing Drive, Block COM1, Singapore 117417, Republic of Singapore*
[†]*leonghw@comp.nus.edu.sg*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

A synteny block represents a set of contiguous genes located within the same chromosome and well conserved among various species. Through long evolutionary processes and genome rearrangement events, large numbers of synteny blocks remain highly conserved across multiple species. Understanding distribution of conserved gene blocks facilitates evolutionary biologists to trace the diversity of life, and it also plays an important role for orthologous gene detection and gene annotation in the genomic era. In this work, we focus on collinear synteny detection in which the order of genes is required and well conserved among multiple species. To achieve this goal, the suffix tree based algorithms for efficiently identifying homologous synteny blocks was proposed. The traditional suffix tree algorithm was modified by considering a chromosome as a string and each gene in a chromosome is encoded as a symbol character. Hence, a suffix tree can be built for different query chromosomes from various species. We can then efficiently search for conserved synteny blocks that are modeled as overlapped contiguous edges in our suffix tree. In addition, we defined a novel Synteny Block Conserved Index (*SBCI*) to evaluate the relationship of synteny block distribution between two species, and which could be applied as an evolutionary indicator for constructing a phylogenetic tree from multiple species instead of performing large computational requirements through whole genome sequence alignment.

*Keywords*: collinear; synteny block; suffix tree; orthologous gene.

## 1. Introduction

The terminology of synteny originally represents gene loci located on the same chromosome[1]. However, it is currently emphasized as a group of genes that are

Co-corresponding authors: [*]Dr. Tun-Wen Pai, Department of Computer Science and Engineering & Center of Excellence for the Oceans, National Taiwan Ocean University, No. 2, Peining Road, Keelung, 20224, Taiwan, R.O.C. TEL: +886-2-24622192 ext. 6618, FAX: +886-2-24623249, E-mail: twp@mail.ntou.edu.tw
[†]Dr. Hon Wai Leong, Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417, TEL: +65-65162903, FAX: +65-67794580, E-mail: leonghw@comp.nus.edu.sg

simultaneously conserved on corresponding chromosomes across different species through speciation and genome rearrangement events without guaranteeing the order of genes. In other words, syntenic genes among various species do not imply the existence of collinear property[2]. Several researches have shown that syntenic genes provide superior features for cross-species comparative genomics[3], such as orthologous gene detection[4,5], gene annotation[6], evolutionary analysis[7] *etc*.. Past research also demonstrated that non-gene regions near a synteny block possessing high probability of acting as regulatory elements. These regulatory elements facilitate biologists to find potential transcription factors within the constraint promoter regions[8,9]. All these previously published reports have shown the importance of synteny characteristics, and several efficient and effective algorithms were continually developed to solve the problems of synteny block detection. These algorithms can be mainly divided into two categories: the first type adopted genome-wide sequence comparison to create conserved anchors initially and applying them to identify syntenic genes, such as Satsuma[10], QUOTA-ALIGN[11], and SyMAP[12]; the second type applied gene annotation and orthologous gene information to retrieve syntenic genes, such as OrthoCluster[13], Cinteny[14], DRIMM-Synteny[15], and i-ADHoRe[16]. The latter methods are computationally more efficient compared to the former. Nevertheless, due to the exponential growth in the generation of genomic data[17], the problem of efficient detection of synteny blocks from multiple genomes is yet to be improved. To achieve this goal, here we have proposed a novel approach by adopting suffix tree based algorithms for efficiently and effectively detecting synteny blocks.

The classical suffix tree algorithm is widely used to solve several problems in computer science such as longest common substring search[18] and data compression[19]. It is also applied in bioinformatics areas such as motif finding[20] and repeat segment identification[21]. The advantage of employing suffix tree algorithm is that once a suffix tree is constructed, the substring search in linear time is achievable. In this study, we considered that the synteny block detection problem is similar to a substring search problem. Therefore, the modified suffix tree algorithm is adopted to efficiently identify synteny blocks from multiple species. To clarify the problem that is solved in this study, we first define a synteny block as a group of genes that are well conserved on corresponding chromosomes across multiple species, and the retrieved genes on a synteny block are required to be contiguous and collinear within a chromosome. Once the corresponding synteny blocks were detected between the query and target genomes, the next problem is how to measure the similarities and levels of conservation in terms of quality and quantity of conserved synteny blocks. Housworth and Postlethwait defined synteny correlation and synteny association which were mainly based on the statistics of orthologous gene numbers[22]. These two approaches could measure the synteny conservation between two species, and were also applied to compare species distance quantitatively. However, from Housworth and Prostlethwait's method, orthologous genes were considered as the basic elements instead of synteny blocks for distance measuring. Therefore, in order to provide a feasible measurement for evaluating the species distance

according to identified synteny blocks, we have designed an intuitive measure based on the number of conserved synteny blocks. In fact, an objective measurement for evaluating synteny block conservation should consider the issues of total number of interspersed synteny blocks (quantity) and number of genes within a conserved synteny block (quality) within a chromosome between two query genomes. Here, we defined an indicator called Synteny Block Conservation Index (*SBCI*) to objectively compare the distances between any two species. This indicator considers both features of richness and evenness of conservation.

In the result section, we have illustrated the performance of our proposed algorithm and compared to Housworth and Postlethwait's method by analyzing 12 representative model species from the Ensembl database[23]. The results have shown that our proposed *SBCI* highly correlated with Housworth and Postlethwait's results. According to the calculated *SBCI*s and constructed phylogenetic tree, we demonstrated that the evolutionary relationships among different organisms could be revealed by adopting *SBCI*s instead of expanding tremendous computational resources in performing huge amounts of genome sequence alignments.

## 2. Materials and Methods

### 2.1. *Ensembl data sets*

A total of 12 genomes of selected model species from Ensembl database (release 69, October 2012) including gene annotation and orthologous information were downloaded for performing initial synteny block analysis. To identify the conserved synteny blocks among various model species, all genes located within an individual chromosome of a specific species were constructed as a single gene ID suffix tree. For verifying phylogenetic tree relationships, the phylogenetic tree from Ensembl was also downloaded for comparison.

### 2.2. *Improved suffix tree algorithms*

In this study, conserved synteny block detection is similar to a pattern searching problem. Each gene ID is considered as a symbol character, while a set of continuous genes is considered as a substring in a context. Hence, all continuous genes within a chromosome are transformed into a gene ID string, and the traditional pattern searching problem by suffix tree algorithm could be applied to conserved synteny block detection problems. The core of the proposed algorithm includes building gene annotation suffix trees (GASTs), searching synteny blocks, and merging duplicate synteny blocks.

The first step is to build gene annotation suffix trees. Unlike general suffix tree algorithms by taking alphabet string as input data, we employed gene annotations and orthologous gene information to construct a gene annotation suffix tree. All genes in a chromosome from a specified model species were transformed into a "gene list" by sorting each annotated gene according to their loci in an individual chromosome. Then, a suffix tree called a gene annotation suffix tree (GAST) was built and stored. In a classical

suffix tree algorithm, a partially identical substring of a suffix should be assigned to the same concatenated edges in the tree. But in the proposed suffix tree algorithm, we adopted orthologous gene information to constrain whether the suffix of concatenated genes should be arranged on the same edges. To simplify the problem in this study, we only applied one-to-one orthologous information. Fig. 1 shows an example of building three GASTs for different species. First, all suffixes from gene lists of three different example species were created and shown in Fig. 1(a). Second, we selected a target species and built its corresponding suffix tree according to the suffixes of the gene list and named as a GAST. An example of a GAST is shown in Fig. 1(b). Third, the second query species of constructed gene lists was built onto the target GAST one-by-one according to their corresponding suffixes and defined by previously collected orthologous information. The result was shown in Fig. 1(c). In the other words, we added a new suffix onto the GAST, the orthologous gene information determined whether the new suffix of genes should be arranged on the same edges. At the last step, the third query species was aligned on the GAST according to its pre-defined suffixes and the results were shown in Fig. 1(d).

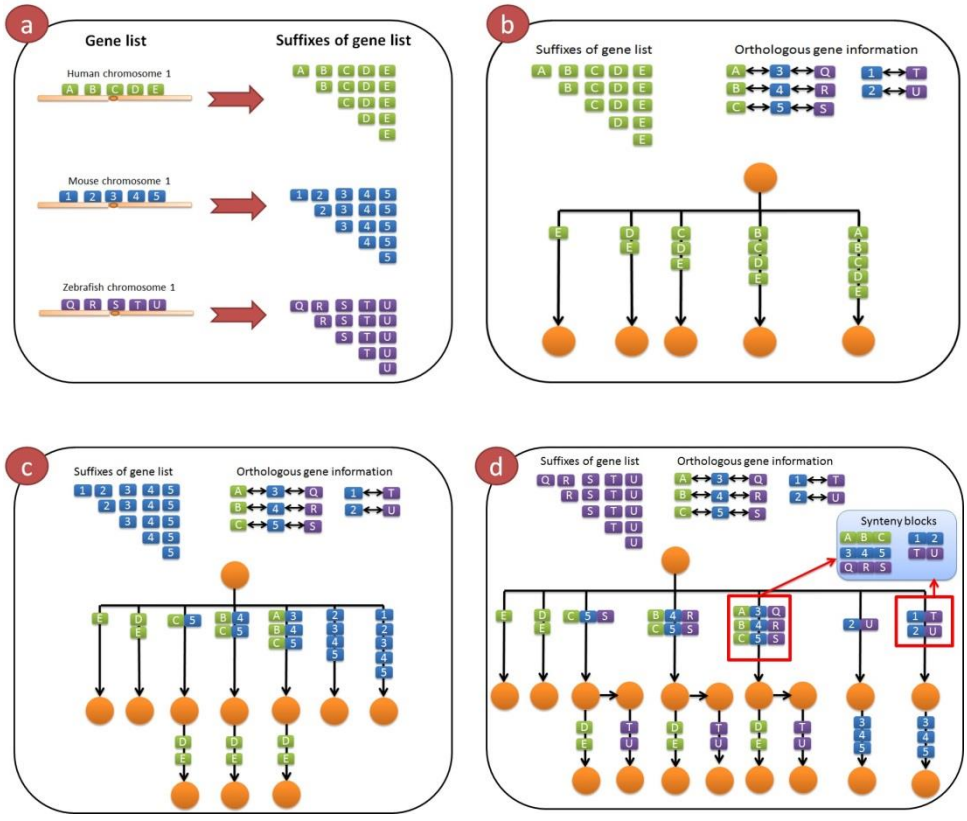



Fig. 1. Example of building a gene annotation suffix tree (GAST) from three different species.

During building a comprehensive GAST, different gene order within two different target species and the intransitivity of orthologous genes would result in different synteny block representations. Fig. 2 is an example to demonstrate intransitivity of orthologous genes. Hence, to overcome this problem, the proposed systems should be assigned with one species as the reference species and the rest of species as the target species before building the GAST. Accordingly, a GAST could be constructed by using the center star approach. In conclusions, a GAST of multiple species would be constructed by taking the reference species as an anchor and the initial GAST should be constructed prior to the rest of the suffixes from all other target species. According to orthologous relationships, the final GAST was built sequentially by checking the orthologous relation between the specified target species and the on-going reference species rather than considering all mutual relationships between any two species pairs.

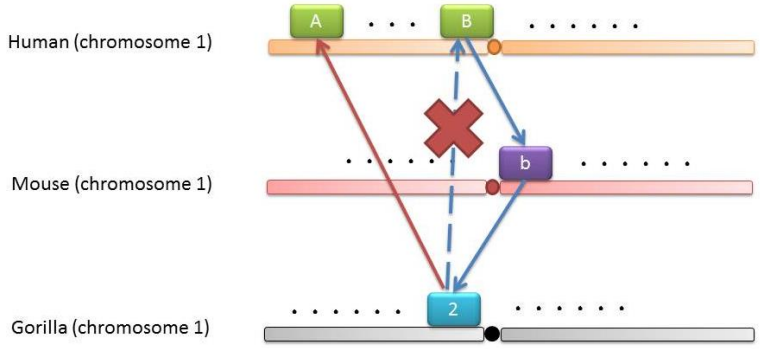



Fig. 2. A transitive problem of defined orthologous genes. The gene "B" of Human chromosome 1 is defined as an orthologous gene of "b" in mouse chromosome 1, and the gene "b" of mouse is also orthologous to the gene "2" in gorilla chromosome 1. However, the gene "2" in gorilla chromosome 1 could not be assumed possessing orthologous characteristics to gene "B" in human chromosome 1. However, the gene "2" in gorilla chromosome 1 might be orthologous with gene "A" in human chromosome 1.

Once a GAST has been constructed, the next step is to detect all synteny blocks on the GAST. It can be efficiently searched for all synteny blocks which are conserved for K species among N species. The K could be selected and ranged from 2 to N. In other words, if we built a GAST containing N species, any synteny block conserved among any 2 to K species could be defined and searched with the same time-complexity in the same GAST. Therefore, different levels of conservation of synteny blocks could be customized by setting different thresholds. In Fig. 1, we could only obtain the synteny block of "(A, B, C), (3, 4, 5), (Q, R, S)" by setting a species threshold of 3. It means that the system will traverse all edges of the GAST and extract edges which were annotated with 3 different species. However, if the threshold was set as 2, the synteny blocks of "(1, 2), (T, U)" and "(A, B, C), (3, 4, 5), (Q, R, S)" will be retrieved for further analysis.

After identifying all synteny blocks from a GAST, the final step is to merge all duplicated synteny blocks. Due to the specific features of a suffix tree structure and different searching thresholds for synteny blocks, the system will obtain duplicated

synteny blocks frequently. Fig. 3 is an example of merging duplicated synteny blocks, of which synteny blocks "(C, 4, c)", "(B, 3, b), (C, 4, c)" and "(A, 1, a), (B, b), (C, c)" in the GAST were obtained by setting a species threshold of 2. According to the overlapped characteristics, these synteny blocks could be merged into one integrated synteny block "(A, 1, a), (B, 3, b), (C, 4, c)". However, there is one gap in the synteny block "(A, 1, a), (B, 3, b), (C, 4, c)", meaning that the proposed algorithm could also retrieve synteny blocks with limited gaps.

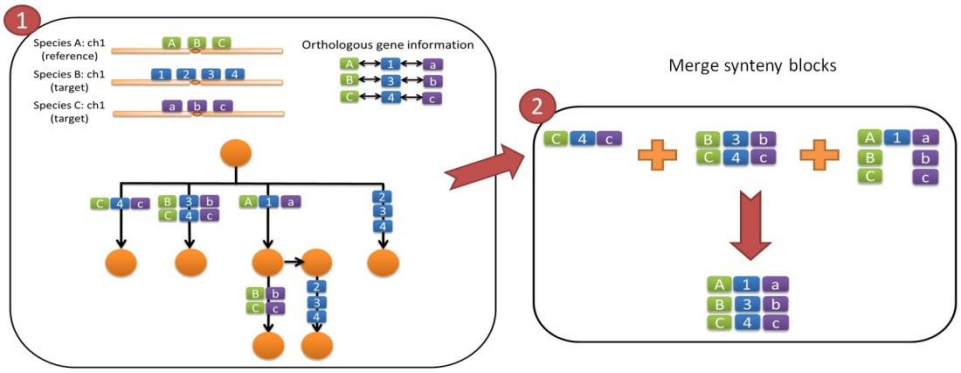



Fig. 3. An example of merging duplicated synteny blocks.

### 2.3. *The features of improved suffix tree*

There are two key features in the proposed suffix tree based synteny block detection algorithm. The first one is that the proposed algorithm allows multiple species comparison for efficient identification of conserved synteny blocks. Users can analyze multiple species simultaneously by building one GAST only. The other benefit offered is the customized settings of identifying synteny block with K conserved species among N species. Users could assign different thresholds to retrieve various conserved levels of synteny blocks and evaluate chromosome evolution in various lineages. Besides, the proposed algorithm is able to retrieve synteny blocks with limited gaps in general.

### 2.4. *Measurement of synteny block conservation*

Synteny block conservation reflects the relative evolutionary rates of chromosomes across genetic lineages. Therefore, a feasible indicator of showing distances between species through the conservation level in synteny block could provide a broad view in phylogenic relationships. Hence, we define a simple but effective indicator of "Synteny Block Conserved Index (*SBCI*)" to evaluate the evolutionary distance between two species regarding the conserved number of synteny blocks. The *SBCI* indicator is composed of two scores: the chromosome dispersion score (*CDS*) and the synteny gene score (*SGS*). The *CDS* is an index of measuring chromosomal rearrangement degree between two species, while the *SGS* evaluates the conserved level between two species.

Unlike the general phylogenetic tree employing sequence alignment approach, the proposed system featured the chromosomal rearrangement factor and conserved levels of syntenic genes to evaluate the phylogenic relationship between two species.

The value of *CDS* ranges from 0 to 1. Higher *CDS* values represent lower chromosome rearrangement degree between the reference species and the target species, and *vice versa*. The *CDS* indicator is defined in the following Eq. (1):

$$CDS(\mathrm{R},\mathrm{T}) = \frac{1}{CN_R}\sum_{\mathrm{i}=1}^{CN_R}\sum_{\mathrm{j}=1}^{CN_T}\left(\lambda\frac{PSB_{ij}}{TNPSB_i} + (1-\lambda)\frac{OG_{ij}}{TNOG_i}\right)^2 \quad (1)$$

$$PSB_{ij} = \sum_{x=1}^{NSB_{ij}}\sum_{y=1}^{C(SBL_x,2)}\left(\frac{\sum_1^{SSBL_y}\begin{cases}1, consistent\ strandedness\\ \beta, inconsistent\ strandedness\end{cases}}{SSBL_y}\right) \quad (2)$$

where $PSB_{ij}$ represents all possible combinations of each conserved synteny block between chromosome "*j*" of the target species and chromosome "*i*" of the reference species; $OG_{ij}$ denotes the number of orthologous genes between chromosome "*j*" of the target species and chromosome "*i*" of the reference species; $TNPSB_i$ is a total number of accumulation of *PSB* from the target species with respect to the chromosome "*i*" of reference species; $TNOG_i$ is a total number of accumulated orthologous genes from the target species with respect to the chromosome "*i*" of the reference species; $CN_T$ is the chromosome number of the target species and $CN_R$ is the chromosome number of the reference species; $\lambda$ is a weighting coefficient for chromosome dispersion assessment between synteny block and orthologous gene conservation. In this study, the default value of $\lambda$ is set as 0.5 for even weights on both factors.

To compare conservation of strandedness of two synteny blocks from two different species, we define $PSB_{ij}$ containing strandedness information as in Eq. (2). For each chromosome pair, $NSB_{i,j}$ represents the number of synteny block between chromosome "*j*" of the target species and chromosome "*i*" of the reference species; $SBL_x$ represents the total gene number of the synteny block "x"; $C(SBL_x, 2)$ denotes all possible 2-combinations of sub-synteny blocks in the current synteny block; $SSBL_y$ indicates the number of genes being included in the sub-synteny block *y*, and $\beta$ is an assigned float value as a score proportional to the level of strandedness similarity. For example, $\beta$ is assigned with a value of 1 which makes the strandedness independent of *CDS*; $\beta$ is assigned with a value of 0 which makes inconsistent strandedness genes negligible for *CDS* calculation. In this study, the default $\beta$ was set as 0.5 for all experiments.

The *SGS*'s is also ranging from 0 to 1. Higher scores indicate higher conservation levels between the reference species and the target species, and *vice versa*. The *SGS* is defined in Eq. (3):

$$SGS = \frac{Uniq\left(\sum_{i=1}^{CN_R}\sum_{j=1}^{CN_T} CSBG_{ij}\right)}{OG_{RT}} \tag{3}$$

where *Uniq*() is a function of removing duplicated genes which belongs to the reference species and calculated more than once at *CSBG* accumulation computing; $CSBG_{ij}$ is the total number of genes in all conserved synteny blocks between chromosome "*j*" from the target species and chromosome "*i*" from the reference species; $OG_{RT}$ is the total orthologous genes between the reference species and the target species.

To combine both *CDS* and *SGS* indicators, an integrated indicator of *SBCI* is calculated according to Eq. (4). The parameters α and (1-α) are weighting coefficients for both chromosome dispersion and synteny conservation features respectively. The default value of α is set as 0.5 for equally weighted assumptions. The final value of an *SBCI* ranges from 0 to 1 as well. A larger *SBCI* value reflects close distance between the reference species and the target species, while a lower value of *SBCI* implies the remote distance between two species. The *SBCI* is defined as the following equation.

$$SBCI = \alpha \frac{CDS(R,T) + CDS(T,R)}{2} + (1-\alpha)SGS. \tag{4}$$

All these indicators between any two model species will be calculated, and selected genomes from Ensembl database will be displayed for discussion.

## 3. Experimental Results

To demonstrate evolutionary relationships among model species, we have selected 12 representative species from different species groups in Ensembl database and calculated mutual *SBCI*s for all species. These 12 representative species include *Homo sapien*s (Human), *Danio rerio* (Zebrafish), *Mus musculus* (Mouse), *Canis familiaris* (Dog), *Gallus gallus* (Chicken), *Monodelphis domestica* (Opossum), *Ciona intestinalis* (*C. intestinalis*), *Caenorhabditis elegans* (*C. elegans*), *Drosophila melanogaster* (Fruitfly), *Saccharomyces cerevisiae* (Yeast), *Ornithorhynchus anatinus* (Platypus), and *Gorilla gorilla* (Gorilla). In order to verify the proposed *SBCI*, we also calculated synteny correlation and synteny association indicators proposed by Housworth and Prostlethwait. These two indicators were also applied to estimate the distances among species. Accordingly, we compared the proposed *SBCI* with Housworth and Prostlethwait's method and try to evaluate whether the results from these three indicators were consistent.

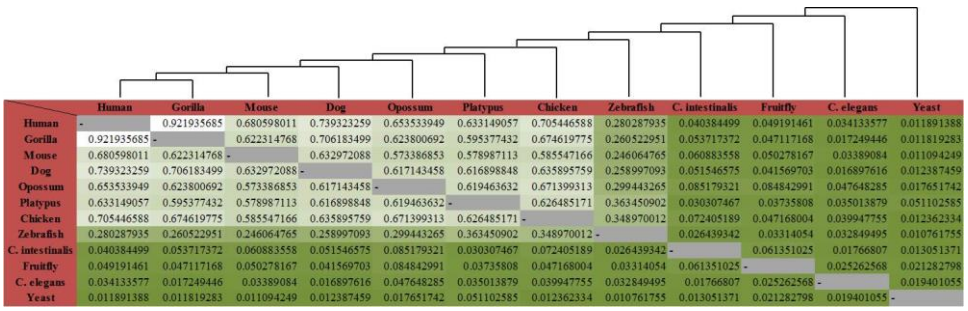
There are in total 66 possible combined pairs from 12 representative species in this study. First, all 66 synteny block matrices were generated. Each matrix of dimension r * c collects the number of conserved synteny blocks in each cell, and the number in the (*i*, *j*) cell represents the number of conserved synteny blocks between $i^{th}$ chromosome in the first species and $j^{th}$ chromosome in the second species. We summed up the number of conserved synteny blocks from all chromosome pairs to obtain the total number of conserved synteny blocks between each species pair, and the mutually conserved synteny

blocks are statistically shown in a newly constructed matrix for all 12 representative species. Furthermore, we also calculate the number of all genes within all detected synteny blocks. Since the both built matrices are symmetric, we integrate both matrices and shown in Table 1. The order of species is sorted by species distance of Ensembl phylogenetic tree by taking Human as the reference species. The upper right triangle of Table 1 is represented as total number of conserved synteny blocks, and the lower left triangle is represented as total number of genes in all synteny blocks.

Table 1. Statistics of mutually conserved synteny blocks for 12 representative species.

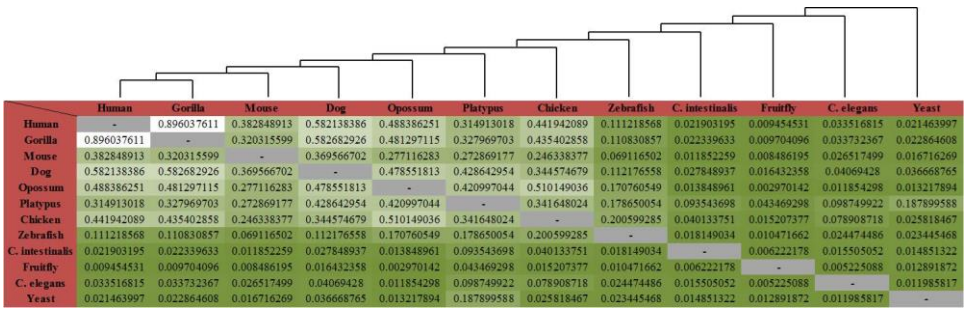
| | **Human** | **Gorilla** | **Mouse** | **Dog** | **Opossum** | **Platypus** | **Chicken** | **Zebrafish** | ***C. intestinalis*** | **Fruitfly** | ***C. elegans*** | **Yeast** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Human** | - | 2942 | 2807 | 3004 | 2858 | 418 | 2542 | 1437 | 5 | 5 | 1 | 0 |
| **Gorilla** | 15428 | - | 3120 | 3174 | 2864 | 397 | 2468 | 1253 | 8 | 4 | 0 | 0 |
| **Mouse** | 14508 | 12871 | - | 3050 | 2924 | 424 | 2594 | 1431 | 7 | 4 | 1 | 0 |
| **Dog** | 13737 | 12248 | 13492 | - | 2897 | 409 | 2518 | 1327 | 8 | 4 | 0 | 0 |
| **Opossum** | 11084 | 9787 | 11069 | 10525 | - | 411 | 2434 | 1283 | 10 | 5 | 1 | 0 |
| **Platypus** | 1294 | 1129 | 1274 | 1201 | 1190 | - | 395 | 179 | 0 | 0 | 0 | 0 |
| **Chicken** | 8877 | 7733 | 8706 | 8180 | 7672 | 1111 | - | 1275 | 7 | 4 | 1 | 0 |
| **Zebrafish** | 3470 | 2905 | 3406 | 3137 | 3023 | 426 | 3074 | - | 2 | 1 | 1 | 0 |
| ***C. intestinalis*** | 10 | 16 | 14 | 16 | 20 | 0 | 14 | 4 | - | 3 | 0 | 0 |
| **Fruitfly** | 10 | 8 | 8 | 8 | 10 | 0 | 8 | 2 | 6 | - | 0 | 0 |
| ***C. elegans*** | 2 | 0 | 2 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | - | 0 |
| **Yeast** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | - |

Once the synteny block matrices were obtained from the suffix tree algorithms, the corresponding *SBCI* values could be obtained immediately by employing Eq. (4) and the results can be applied to construct a similarity matrix for representing the relationship among various species. For comparison, the species relationships described by various approaches including *SBCI*, synteny correlation, and synteny association indicators were constructed and shown in Fig 4 to 6. On top of each matrix represents species distance relationship from Ensembl phylogenetic tree. To clearly represent the distance between two species, a heat map format is applied to visualize the relationship. The lighter colors represent higher similarities between two correlated species, and the darker colors for distantly related species. From Fig. 4 to 6, it can be observed that color shades from upper-left to lower-right corners appeared from light to dark color shades. It reflects that the relationships calculated by *SBCI* indicators are highly consistent with synteny correlation and association indicators. The Pearson Correlation Coefficient between *SBCI* and synteny correlation indicator is 0.94, and for *SBCI* and synteny association is 0.95. The results showed consistent relationship between the proposed indicator and previously known indicators.



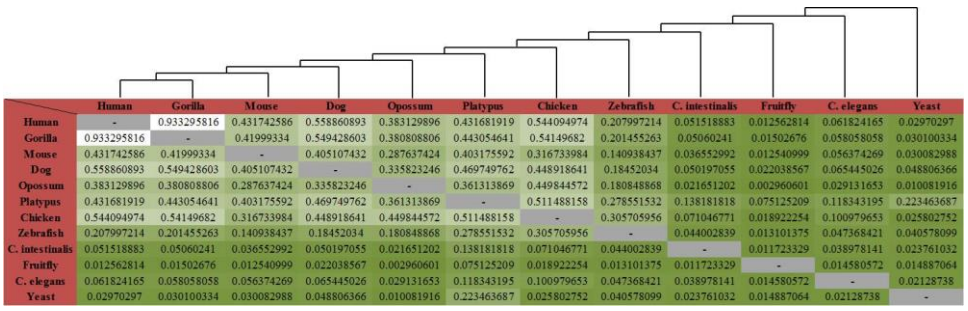
| | Human | Gorilla | Mouse | Dog | Opossum | Platypus | Chicken | Zebrafish | C. intestinalis | Fruitfly | C. elegans | Yeast |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | - | 0.921935685 | 0.680598011 | 0.739323259 | 0.653533949 | 0.633149057 | 0.705446588 | 0.280287935 | 0.040384499 | 0.049191461 | 0.034133577 | 0.011891388 |
| Gorilla | 0.921935685 | - | 0.622314768 | 0.706183499 | 0.623800692 | 0.595377432 | 0.674619775 | 0.260522951 | 0.053717372 | 0.047117168 | 0.017249446 | 0.011819283 |
| Mouse | 0.680598011 | 0.622314768 | - | 0.632972088 | 0.573386853 | 0.578987113 | 0.585547166 | 0.246064765 | 0.060883558 | 0.050278167 | 0.03389084 | 0.011094249 |
| Dog | 0.739323259 | 0.706183499 | 0.632972088 | - | 0.617143458 | 0.616898848 | 0.635895759 | 0.258997093 | 0.051546575 | 0.041569703 | 0.016897616 | 0.012387459 |
| Opossum | 0.653533949 | 0.623800692 | 0.573386853 | 0.617143458 | - | 0.619463632 | 0.671399313 | 0.299443265 | 0.085179321 | 0.084842991 | 0.047648285 | 0.017651742 |
| Platypus | 0.633149057 | 0.595377432 | 0.578987113 | 0.616898848 | 0.619463632 | - | 0.626485171 | 0.363450902 | 0.030307467 | 0.03735808 | 0.035013879 | 0.051102585 |
| Chicken | 0.705446588 | 0.674619775 | 0.585547166 | 0.635895759 | 0.671399313 | 0.626485171 | - | 0.348970012 | 0.072405189 | 0.047168004 | 0.039947755 | 0.012362334 |
| Zebrafish | 0.280287935 | 0.260522951 | 0.246064765 | 0.258997093 | 0.299443265 | 0.363450902 | 0.348970012 | - | 0.026439342 | 0.03314054 | 0.032849495 | 0.010761755 |
| C. intestinalis | 0.040384499 | 0.053717372 | 0.060883558 | 0.051546575 | 0.085179321 | 0.030307467 | 0.072405189 | 0.026439342 | - | 0.061351025 | 0.01766807 | 0.013051371 |
| Fruitfly | 0.049191461 | 0.047117168 | 0.050278167 | 0.041569703 | 0.084842991 | 0.03735808 | 0.047168004 | 0.03314054 | 0.061351025 | - | 0.025262568 | 0.021282798 |
| C. elegans | 0.034133577 | 0.017249446 | 0.03389084 | 0.016897616 | 0.047648285 | 0.035013879 | 0.039947755 | 0.032849495 | 0.01766807 | 0.025262568 | - | 0.019401055 |
| Yeast | 0.011891388 | 0.011819283 | 0.011094249 | 0.012387459 | 0.017651742 | 0.051102585 | 0.012362334 | 0.010761755 | 0.013051371 | 0.021282798 | 0.019401055 | - |

Fig. 4. A mutual heat map relationship among 12 representative species obtained by employing *SBCI* indicators. Both weighting coefficients of λ and α were set as the default values of 0.5. Light colors and dark color represent high similarity and low similarity between two species respectively.



| | Human | Gorilla | Mouse | Dog | Opossum | Platypus | Chicken | Zebrafish | C. intestinalis | Fruitfly | C. elegans | Yeast |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | - | 0.896037611 | 0.382848913 | 0.582138386 | 0.488386251 | 0.314913018 | 0.441942089 | 0.111218568 | 0.021903195 | 0.009454531 | 0.033516815 | 0.021463997 |
| Gorilla | 0.896037611 | - | 0.320315599 | 0.582682926 | 0.481297115 | 0.327969703 | 0.435402858 | 0.110830857 | 0.022339633 | 0.009704096 | 0.033732367 | 0.022864608 |
| Mouse | 0.382848913 | 0.320315599 | - | 0.369566702 | 0.277116283 | 0.272869177 | 0.246338377 | 0.069116502 | 0.011852259 | 0.008486195 | 0.026517499 | 0.016716269 |
| Dog | 0.582138386 | 0.582682926 | 0.369566702 | - | 0.478551813 | 0.428642954 | 0.344574679 | 0.112176558 | 0.027848937 | 0.016432358 | 0.04069428 | 0.036668765 |
| Opossum | 0.488386251 | 0.481297115 | 0.277116283 | 0.478551813 | - | 0.420997044 | 0.510149036 | 0.170760549 | 0.013848961 | 0.002970142 | 0.011854298 | 0.013217894 |
| Platypus | 0.314913018 | 0.327969703 | 0.272869177 | 0.428642954 | 0.420997044 | - | 0.341648024 | 0.178650054 | 0.093543698 | 0.043469298 | 0.098749922 | 0.187899588 |
| Chicken | 0.441942089 | 0.435402858 | 0.246338377 | 0.344574679 | 0.510149036 | 0.341648024 | - | 0.200599285 | 0.040133751 | 0.015207377 | 0.078908718 | 0.025818467 |
| Zebrafish | 0.111218568 | 0.110830857 | 0.069116502 | 0.112176558 | 0.170760549 | 0.178650054 | 0.200599285 | - | 0.018149034 | 0.010471662 | 0.024474486 | 0.023445468 |
| C. intestinalis | 0.021903195 | 0.022339633 | 0.011852259 | 0.027848937 | 0.013848961 | 0.093543698 | 0.040133751 | 0.018149034 | - | 0.006222178 | 0.015505052 | 0.014851322 |
| Fruitfly | 0.009454531 | 0.009704096 | 0.008486195 | 0.016432358 | 0.002970142 | 0.043469298 | 0.015207377 | 0.010471662 | 0.006222178 | - | 0.005225088 | 0.012891872 |
| C. elegans | 0.033516815 | 0.033732367 | 0.026517499 | 0.04069428 | 0.011854298 | 0.098749922 | 0.078908718 | 0.024474486 | 0.015505052 | 0.005225088 | - | 0.011985817 |
| Yeast | 0.021463997 | 0.022864608 | 0.016716269 | 0.036668765 | 0.013217894 | 0.187899588 | 0.025818467 | 0.023445468 | 0.014851322 | 0.012891872 | 0.011985817 | - |

Fig. 5. A mutual heat map of relationship among 12 representative species obtained by employing synteny correlation indicators. Light colors and dark color represent high and low similarity respectively between two species.



| | Human | Gorilla | Mouse | Dog | Opossum | Platypus | Chicken | Zebrafish | C. intestinalis | Fruitfly | C. elegans | Yeast |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | - | 0.933295816 | 0.431742586 | 0.558860893 | 0.383129896 | 0.431681919 | 0.544094974 | 0.207997214 | 0.051518883 | 0.012562814 | 0.061824165 | 0.02970297 |
| Gorilla | 0.933295816 | - | 0.41999334 | 0.549428603 | 0.380808806 | 0.443054641 | 0.54149682 | 0.201455263 | 0.05060241 | 0.01502676 | 0.058058058 | 0.030100334 |
| Mouse | 0.431742586 | 0.41999334 | - | 0.405107432 | 0.287637424 | 0.403175592 | 0.316733984 | 0.140938437 | 0.036552992 | 0.012540999 | 0.056374269 | 0.030082988 |
| Dog | 0.558860893 | 0.549428603 | 0.405107432 | - | 0.335823246 | 0.469749762 | 0.448918641 | 0.18452034 | 0.050197055 | 0.022038567 | 0.065445026 | 0.048806366 |
| Opossum | 0.383129896 | 0.380808806 | 0.287637424 | 0.335823246 | - | 0.361313869 | 0.449844572 | 0.180848868 | 0.021651202 | 0.002960601 | 0.029131653 | 0.010081916 |
| Platypus | 0.431681919 | 0.443054641 | 0.403175592 | 0.469749762 | 0.361313869 | - | 0.511488158 | 0.278551532 | 0.138181818 | 0.075125209 | 0.118343195 | 0.223463687 |
| Chicken | 0.544094974 | 0.54149682 | 0.316733984 | 0.448918641 | 0.449844572 | 0.511488158 | - | 0.305705956 | 0.071046771 | 0.018922254 | 0.100979653 | 0.025802752 |
| Zebrafish | 0.207997214 | 0.201455263 | 0.140938437 | 0.18452034 | 0.180848868 | 0.278551532 | 0.305705956 | - | 0.044002839 | 0.013101375 | 0.047368421 | 0.040578099 |
| C. intestinalis | 0.051518883 | 0.05060241 | 0.036552992 | 0.050197055 | 0.021651202 | 0.138181818 | 0.071046771 | 0.044002839 | - | 0.011723329 | 0.038978141 | 0.023761032 |
| Fruitfly | 0.012562814 | 0.01502676 | 0.012540999 | 0.022038567 | 0.002960601 | 0.075125209 | 0.018922254 | 0.013101375 | 0.011723329 | - | 0.014580572 | 0.014887064 |
| C. elegans | 0.061824165 | 0.058058058 | 0.056374269 | 0.065445026 | 0.029131653 | 0.118343195 | 0.100979653 | 0.047368421 | 0.038978141 | 0.014580572 | - | 0.02128738 |
| Yeast | 0.02970297 | 0.030100334 | 0.030082988 | 0.048806366 | 0.010081916 | 0.223463687 | 0.025802752 | 0.040578099 | 0.023761032 | 0.014887064 | 0.02128738 | - |

Fig. 6. A mutual heat map of relationship among 12 representative species obtained by employing synteny association indicators. Light colors and dark color represent high and low similarity respectively between two species.

## 4. Conclusion

In this paper, we proposed a newly modified suffix tree algorithm to identify conserved synteny blocks among species. The proposed suffix tree algorithm combined with homologous information supports fast search on synteny blocks across multiple species compared to sequence alignment approaches. We also defined a novel indicator of *SBCI* integrating two characteristics of chromosome dispersion condition and conserved synteny gene number to measure the distance between any two species. Compared to the general phylogenetic tree approach which is obtained by taking sequence similarity on whole genome sequences, our proposed system provides a more efficient and effective approach for analyzing phylogenetic relationships regarding the chromosomal rearrangement issue and conservation issue to evaluate the distance between species. In addition, in order to verify the effectiveness of *SBCI* indicator, we compared the proposed *SBCI* indicator with previously published synteny correlation and synteny association indicators. The results have shown consistent relationship among these indicators. For the modified suffix tree algorithms, there are several advantages compared to existing methods. For example, only one GAST should be constructed for multiple species comparison. Furthermore, once a GAST is constructed, all synteny blocks with K conserved species among N species could be obtained immediately. Another benefit is that the constructed GASTs could be stored and applied to any newly sequenced model species. However, the proposed suffix tree algorithm based approach for identifying conserved synteny blocks still possesses some disadvantages such as gap tolerance problems. Due to this limitation on the suffix tree structure itself, we could only detect synteny blocks with limited gaps instead of comprehensive gap tolerance. Although a synteny block is strictly defined with contiguous and collinear characteristics, a synteny block with gaps should be considered for practical biology considerations. The other disadvantage is that we only applied one-to-one orthologous relationship to construct a GAST in this study. Many-to-many orthologous mapping could complicate the tree building and searching processes for identifying conserved synteny blocks among multiple species. Though there are several algorithms and measurements have been proposed for synteny gene related topics, here we proposed an alternative approach for efficient searching algorithm and an intuitive measuring indicator based on conserved synteny blocks. We believe that the obtained and previously stored conserved synteny blocks between any species pair could be easy to be updated and employed to evaluate the evolutionary relationships under any challenging environment.

## Acknowledgments

This work is supported by the Center of Excellence for the Oceans, National Taiwan Ocean University and National Science Council, Taiwan, R.O.C. (NSC 102-2321-B-019 - 001 to Tun-Wen Pai)

## References


1 E. Passarge, B. Horsthemke, and R. A. Farber, "Incorrect use of the term synteny," *Nature Genetics,* vol. 23, pp. 387-387, Dec 1999.

2 H. Tang, J. E. Bowers, X. Wang, R. Ming, M. Alam, and A. H. Paterson, "Synteny and collinearity in plant genomes," *Science,* vol. 320, pp. 486-8, Apr 25 2008.

3 M. T. Webster, N. G. Smith, M. J. Lercher, and H. Ellegren, "Gene expression, synteny, and local similarity in human noncoding mutation rates," *Mol Biol Evol,* vol. 21, pp. 1820-30, Oct 2004.

4 J. Jun, I. I. Mandoiu, and C. E. Nelson, "Identification of mammalian orthologs using local synteny," *BMC Genomics,* vol. 10, Dec 23 2009.

5 M. Zhang and H. W. Leong, "BBH-LS: an algorithm for computing positional homologs using sequence and gene context similarity," *BMC Syst Biol,* vol. 6 Suppl 1, p. S22, Jul 16 2012.

6 A. P. Yelton, B. C. Thomas, S. L. Simmons, P. Wilmes, A. Zemla, M. P. Thelen, N. Justice, and J. F. Banfield, "A Semi-Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and Poorly Annotated Bacterial and Archaeal Genes," *Plos Computational Biology,* vol. 7, Oct 2011.

7 C. Kemkemer, M. Kohn, D. N. Cooper, L. Froenicke, J. Hogel, H. Hameister, and H. Kehrer-Sawatzki, "Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution," *Bmc Evolutionary Biology,* vol. 9, Apr 24 2009.

8 D. K. Goode, P. Snell, S. F. Smith, J. E. Cooke, and G. Elgar, "Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3," *Genomics,* vol. 86, pp. 172-181, Aug 2005.

9 A. P. Lee, E. G. L. Koh, A. Tay, S. Brenner, and B. Venkatesht, "Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 103, pp. 6994-6999, May 2 2006.

10 M. G. Grabherr, P. Russell, M. Meyer, E. Mauceli, J. Alfoldi, F. Di Palma, and K. Lindblad-Toh, "Genome-wide synteny through highly sensitive sequence alignment: Satsuma," *Bioinformatics,* vol. 26, pp. 1145-1151, May 1 2010.

11 H. B. Tang, E. Lyons, B. Pedersen, J. C. Schnable, A. H. Paterson, and M. Freeling, "Screening synteny blocks in pairwise genome comparisons through integer programming," *BMC Bioinformatics,* vol. 12, Apr 18 2011.

12 C. Soderlund, M. Bomhoff, and W. M. Nelson, "SyMAP v3.4: a turnkey synteny system with application to plant genomes," *Nucleic Acids Res,* vol. 39, May 2011.

13 X. Zeng, M. J. Nesbitt, J. Pei, K. Wang, I. A. Vergara, and N. Chen, "OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics," presented at the Proceedings of the 11th international conference on Extending database technology: Advances in database technology, Nantes, France, 2008.

14 A. U. Sinha and J. Meller, "Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms," *BMC Bioinformatics,* vol. 8, p. 82, 2007.

15 S. K. Pham and P. A. Pevzner, "DRIMM-Synteny: decomposing genomes into evolutionary conserved segments," *Bioinformatics*, vol. 26, pp. 2509-16, Oct 15 2010.
16 S. Proost, J. Fostier, D. De Witte, B. Dhoedt, P. Demeester, Y. Van de Peer, and K. Vandepoele, "i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets," *Nucleic Acids Res*, vol. 40, p. e11, Jan 2012.
17 M. Baker, "Next-generation sequencing: adjusting to data overload," *Nature Methods*, vol. 7, pp. 495-499, Jul 2010.
18 D. Gusfield, *Algorithms on strings, trees, and sequences: computer science and computational biology*: Cambridge University Press, 1997.
19 N. J. Larsson, "Extended application of suffix trees to data compression," in *Data Compression Conference, 1996. DCC '96. Proceedings*, 1996, pp. 190-199.
20 J. Nicolas, P. Durand, G. Ranchy, S. Tempel, and A. S. Valin, "Suffix-tree analyser (STAN): looking for nucleotidic and peptidic patterns in chromosomes," *Bioinformatics*, vol. 21, pp. 4408-10, Dec 15 2005.
21 S. Kurtz and C. Schleiermacher, "REPuter: fast computation of maximal repeats in complete genomes," *Bioinformatics*, vol. 15, pp. 426-7, May 1999.
22 E. A. Housworth and J. Postlethwait, "Measures of synteny conservation between species pairs," *Genetics*, vol. 162, pp. 441-448, Sep 2002.
23 P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kahari, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. Searle, "Ensembl 2013," *Nucleic Acids Res*, vol. 41, pp. D48-55, Jan 2013.

**Yu-Lun Chen** received his B.S.E. degree in Electrical Engineering, from Tamkang University, Taiwan, in 2011. He is currently a Master student at National Taiwan Ocean University, Keelung, Taiwan. His research interests include bioinformatics and cloud computing.

**Chien-Ming Chen** received the M.S.E. degree in computer science and engineering from National Taiwan Ocean University in 2008. He is currently a Ph.D. student in the same university. His research interests include bioinformatics and information retrieval.

**Tun-Wen Pai** received the degree in Electrical Engineering from National Taipei Institute of Technology, Taipei, Taiwan, in 1984, the M.S.E. degree in Electrical and Computer Engineering from Johns Hopkins University, Baltimore, MD, in 1989, and the Ph.D. degree in Electrical and Computer Engineering from Duke University, Durham, NC, in 1993. He is currently a Professor of Dept. of Computer Science and Engineering at National Taiwan Ocean University, Keelung, Taiwan. His research interests are in algorithms and system design for multimedia modeling and biological data analysis.

**Hon Wai Leong** is an Associate Professor at the Department of Computer Science at the National University of Singapore. He received his B.Sc. (Hon) degree in Mathematics from the University of Malaya and the Ph.D. degree in Computer Science from the University of Illinois at Urbana-Champaign. His research interest is in the design of optimization algorithms for problems from diverse application areas including VLSI-CAD, transportation logistics, multimedia systems and computational biology. In computational biology, his current interests includes computational proteomics, fragment assembly, comparative genomics and analysis of PPI networks. He has a passion for nurturing young talent and conducts many workshops on creative problem solving and computational thinking. In 1992, he started the Singapore training program for the IOI (International Olympiad in Informatics). He is a member of ACM, IEEE, ISCB and a Fellow of the Singapore Computer Society. (homepage: http://www.comp.nus.edu.sg/_leonghw/.)

**Ket Fah Chong** is currently a teaching assistant at the National University of Singapore (NUS), School of Computing. He received his B.Sc (Hon), M.Sc and Ph.D. degrees in Computing from NUS. His research interests include algorithms, side-chain conformation prediction and protein sequencing by mass spectrometry.