# Hidden Markov Model Based Graph Construction Process for DNA Sequence Assembly



# Hidden Markov Model Based Graph Construction Process for DNA Sequence Assembly

**Xia Zhang[1], Weimin Qi[1] and Zhiming Zhan[1,*]**

[1]School of Physics and Information Engineering, Jianghan University, Wuhan, China, 430056

*Corresponding author e-mail: jasonzzm@tom.com

**Abstract**. Through using next-generation sequencers to decode DNA symbols has been a majorly breakthrough in the area of genomic research for decades. A plenty of current approaches of next-generation sequencers with high throughput rates as well as relatively low costs, but it is still challenged for the assembly of the reads which those sequencers produces. We proposed, in this paper, a novel Hidden Markov Model based (HMM-based) approach for next-generation genome sequence assembly programs. The paper introduces the major challenges that currently existed assemblers encounter in the next-generation environment, and four basic stages included in our proposed method: a) pre-processing filtering, b) a graph construction process, c) a graph simplification process, d) post-processing filtering. Experimental results prove the performance of the new approach meets or exceeds the state-of-art by testing a number of DNA open-source datasets.

## 1. Introduction

With the advent of massively parallel sequencing technologies, biological research has rapidly changed recently, in one of which approach is well known as next-generation sequencing (NGS) [1]. High throughputs at low costs are its notable performance [2, 3] for those sequences for short lengths.

The difficulty of genome assembly is the impossibility of directly sequencing the whole genomic sequence within one read via using any of current genome sequencing methods [4]. The shotgun for sequencing techniques is to divide a whole genome sequence into a number of random reads and then to independently sequence each of reads [5], all of which will be followed by the process – called as genome assembly – of the reconstruction of a whole genome via assembling those reads together [6] back up to the chromosomal level. Sanger method, during the past two decade, was the top approach in genomic sequencing, which works for long reads (800-1000 base pairs) and outputs low throughput with high costs [1, 4, 7]. The development of a descent framework which organizes the procedure of establishing an assembler being a pipeline with interleaved processes [8, 9] will be the first step to overcome the assembly challenge in NGS. NGS assembly processes, generally, contained four stages: pre-processing filtering, graph construction, graph simplification, and post-processing filtering [10].

A large amount of communication information should be in transferring among above four stages in which of each stage only work for its individual input for the purpose of producing the outputs which can reach to be maximally functional by itself [11]. These stages can be found in a majority of recent assemblers in the next-generation environment [12]; however, some other assemblers postpone the pre-processing filtering until the later stages [13]. To propose an assembler aiming to the next-

generation has a plenty of challenges - for instances, the sequencers of high-throughput nature [14], sequencing errors [15], short-read lengths [16, 17], and genomic repeats [18], all of which will cause the genome assembly task more complex and complicated and increase the needs for computational resources in hardware and software [19].

## 2. Proposed Approaches and Models

### *2.1. Hidden Markov Model*

Hidden Markov model (HMM) is a statistical or stochastic Markov model working for the system is modeled under the assumption that the modelling is a Markov process with unobserved states. One of the simplest representatives of HMM is dynamic Bayesian network, which is associated with a previous research about the optimal non-linear filtering problem [20] which, as Figure 1 shows, was the first work to define and determine the so-call "forward-backward" procedure.

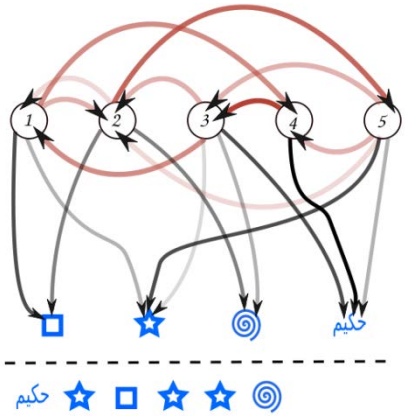


**Figure 1.** Hidden Markov Model with Stochastic Features

A HMM is a model where the hidden variables who control if the mixture component is selected or not based on every observation from data level, are associated with each other via a Markov process instead of independent [21]. During recent years, HMMs have been in generalization as two types – pairwise hidden Markov models and triplet hide Markov models, the latter of which is supposed to handle more complex data structures especially for modelling non-stationary data. Current research claims HMMs may combine with statistical or stochastic significances such as probability for better performance. HMMs together with Viterbi algorithm are applied to estimate the relevance for an output sequence based on a hypothesis, where the statistical significances demonstrate the false positive ratio related to the failure of rejecting the hypothesis for the particular sequence.

### *2.2. Graph Construction Models*

Overlap-based graph construction and K-Spectrum–Based graph construction are the two major graph construction models currently, both of which are widely used in geometric sequencing research.

Overlap-based assemblers (example of Figure 2) start with detecting the overlaps based on a large number of unassembled reads. Secondly, the overlapping information will be put into a construction graph, whose nodes are reads and whose edges are the overlaps between pairwise nodes.

K-Spectrum–Based Construction is another approach in current research. Such assemblers (example of Figure 3) firstly extract all k-mers in the reads, referring to their k-spectrum, where every node is a k-mer in the graph while every edge is k–1 overlapping between pairs of nodes. Generally, an Eulerian path will traversal every edge exactly once in the construction graph which represents the entire chromosome, if the length of sequence is known.

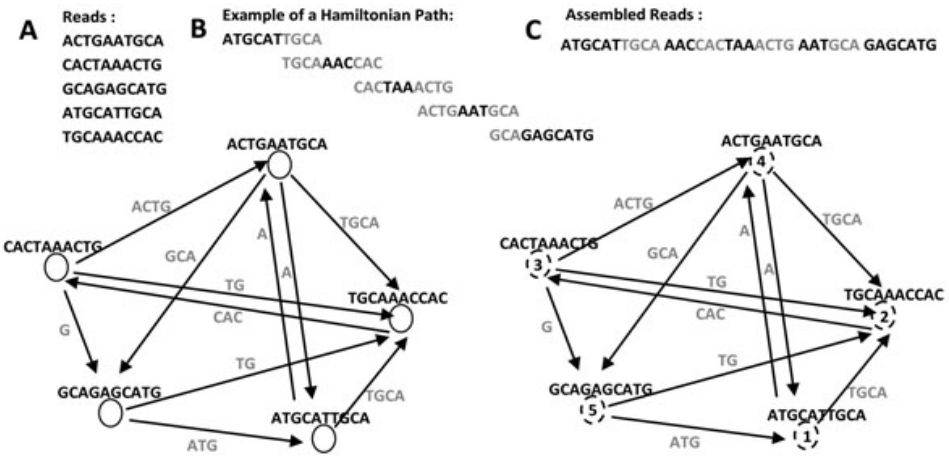


**Figure 2.** Overlap-based Graph Sequencing Construction
(A) Graph nodes are reads, in edges are overlapping between which.
(B) A Hamiltonian path as example traversing every node only once.
(C) Assembled reads based on the above Hamiltonian path in order.

In Figure 2, the layout step aims to find out a shortest Hamiltonian path which traversal each graph node once and exactly once. Thus, a Hamiltonian path is a representative of one assembly solution. Last, the overlaps between the nodes will be merged for construction purpose at the consensus step.

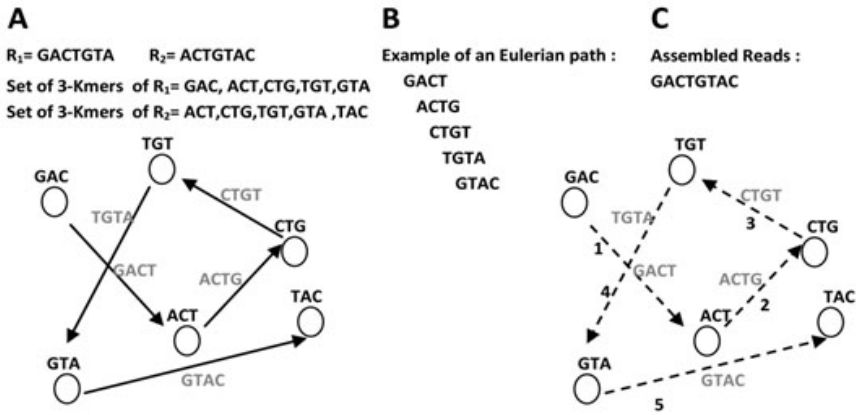


**Figure 3.** K-spectrum–based Graph Sequencing Construction
(A) Nodes are k-mers, whose edges are k–1 overlapping between.
(B) Example showing an Eulerian path traversing every edge only once.
(C) Assembled reads according to the Eulerian path above.

For the case of assembling a genome with high-coverage or high-error profiles, the approach increases the number of both repeated and distinct k-mers in such graph, causing performance downgraded a lot.

## 3. Testing and Results

### *3.1. Datasets*

We have downloaded 38 conserved DNA sequences of E.coli helix-turn helix-5 structure from NBCI conserved domain databases. The multiple sequence alignment process was done by online tool MUSCLE. Next, we extracted the first 37 sequences as the training data set for building HMM profile and left the last one as testing. The HMM profile was made by the following steps:

a) With the multiple sequence alignment results, we first define each column to mutation column or insertion column, by the criteria that in the column if the proportion of gap '-' in that column is larger than 50%. Then we define it as insertion column and otherwise define it as mutation column.

b) In order to handle the over-fitting problem, the pseudo-counts were added differently between mutation or insertion columns. In each mutation column, we add up to 1 for each symbol of A, T, C ,G and '-'. And we add a pseudo insertion state by adding 1 for letter symbol and others are all set to '-'. The positions of adding pseudo insertion states are that at each mutation state if the mutation state in the original multiple sequence alignment results is not followed by an insertion. Thus, we have 141 Mutation state each followed by 141 Insertion state, like M0, I0, M1, I1, …, M141, I141. Here, we assume the emission probability of A, T, C and G in each Insertion state as 0.25, since that in the random case we should observe the probability of A, T, C and G occurring as insertion. Another way is counting the A, T, C and G probability at each Insertion state and add some pseudo counts to get the emission probability for Insertion state. However, it is really possible that our HMM profile may have some bias to represent the true population. Thus, we think 0.25 emission probability for each letter in Insertion State is better.

c) After adding all the pseudo counts as described above, we have the HMM profile. Then, all the transition probability and emission probability are calculated and stored in the transition probability matrix and the emission probability matrix, separately. For Deletion state there is no corresponding emission probability.

### *3.2. Bioinformatics-based HMM Modelling Algorithm*

Based on the HMM (as Figure 4) and its HMM-Profile, after calculating the transition and emission probability matrix, we exploit Dynamic Programming (DP) to do sequencing.

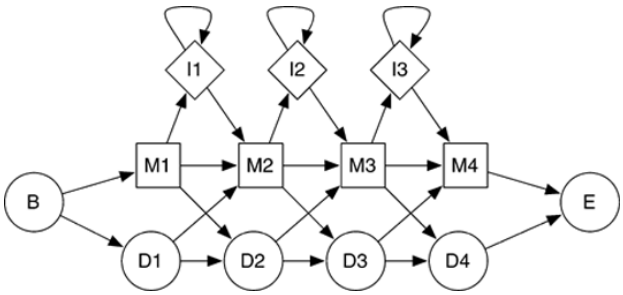


**Figure 4.** Topology Diagram of Hidden Markov Model for DNA Sequence Assembly

The transition probability and its DP recursive function from state I to other states are described as (1) respectively, in which A is the average length of the sequence, and Δ is the pseudo count equaling to 1.

$$\begin{cases} q(I_k, I_k) = \dfrac{A-1}{A} \\ q(I_k, D_{k+1}) = \dfrac{(n - cnt\_letter\_M(k+1)) + \Delta}{n + 2\Delta}, \; V_{i,j}^{M} = \max \begin{Bmatrix} V_{i-1,j-1}^{M} \cdot q(M_{j-1}, M_j) \\ V_{i-1,j-1}^{I} \cdot q(I_{j-1}, M_j) \\ V_{i-1,j-1}^{D} \cdot q(D_{j-1}, M_j) \end{Bmatrix} \cdot p(M_j, x_i) \\ q(I_k, M_{k+1}) = \dfrac{cnt\_letter\_M(k+1) + \Delta}{n + 2\Delta} \end{cases} \quad (1)$$

The transition probability from state M to other states and its DP recursive function are as (2) describes respectively, in which of the terms of A and Δ the same as (1) are.

$$\begin{cases} q(M_k, I_k) = \dfrac{cnt\_letter\_I(k+1) + \Delta}{n + 2\Delta} \\ q(M_k, D_{k+1}) = (1 - q(M_k, I_k)) \cdot \dfrac{(n - cnt\_letter\_M(k+1)) + \Delta}{n + 2\Delta}, \; V_{i,j}^{I} = \max \begin{Bmatrix} V_{i-1,j}^{M} \cdot q(M_j, I_j) \\ V_{i-1,j}^{I} \cdot q(I_j, I_j) \\ V_{i-1,j}^{D} \cdot q(D_j, I_j) \end{Bmatrix} \\ q(M_k, M_{k+1}) = (1 - q(M_k, I_k)) \cdot \dfrac{cnt\_letter\_M(k+1) + \Delta}{n + 2\Delta} \end{cases} \quad (2)$$

The transition probability from state D to other states and its DP recursive function are as (3) shows respectively, where the terms of A and Δ are the same as (1).

$$\begin{cases} q(D_k, I_k) = \dfrac{cnt\_letter\_I(k+1)+\Delta}{n+2\Delta} \\ q(D_k, D_{k+1}) = (1-q(D_k, I_k)) \cdot \dfrac{(n-cnt\_letter\_M(k+1))+\Delta}{n+2\Delta} \\ q(D_k, D_{k+1}) = (1-q(D_k, I_k)) \cdot \dfrac{cnt\_letter\_M(k+1)+\Delta}{n+2\Delta} \end{cases}, \quad V_{i,j}^{D} = \max \begin{Bmatrix} V_{i,j-1}^{M} \cdot q(M_{j-1}, D_j) \\ V_{i,j-1}^{I} \cdot q(I_{j-1}, D_j) \\ V_{i,j-1}^{D} \cdot q(D_{j-1}, D_j) \end{Bmatrix} \quad (3)$$

The base cases involved in the above DP recursion are: $V_{i,\mathrm{j}}^{S} = \begin{cases} 1 & , \quad S = M \quad and \quad i = j \\ 0 & , \quad else \end{cases}$.

We use Viterbi Algorithm to decoding test sequences [22], where the logarithm was applied in calculating each transition or emission probability [23], since a large number of very small real numbers' multiplication could cause computing underflow problem [24].

### *3.3. Experimental Results*

DNA sequences were inputted into the Viterbi decoding algorithm as observation and use it to test our Hidden Markov Model. Since the multiple sequence alignment result for the DNA are known already, we can statistically compare them. The accordance map results are summarized in Table 1.

**Table 1.** HMM Validation Results

| Dataset | Sequence Length | Mis-match | Indels | Matches |
|---|---|---|---|---|
| 1 | 148 | 49 (33.1%) | 15 (10.1%) | 84 (56.8%) |
| 2 | 189 | 58 (30.7%) | 18 (9.5%) | 113 (59.8%) |
| 3 | 202 | 61 (30.2%) | 21 (10.4%) | 120 (59.4%) |
| 4 | 216 | 69 (31.9%) | 22 (10.2%) | 125 (57.9%) |

The result indicates our HMM could accurately decode more than half percent of the sequence. But, the mismatch rate is kind of higher than our expectation. The reason for this could be that the HMM profile sequence (training dataset) itself is not so conserved, since each of the sequence in training dataset is relatively long with 141 mutation state, it is possible that less than half of the mutation state is really conserved across 38 helix-turn-helix-5 structural sequences found in NCBI database. In terms of Indels, our HMM has an error rate around 10%, which is tolerable. Probably, we could adjust our transition probability matrix by changing pseudo counts to increase the accuracy for predicting Indels.

## 4. Future Work

Further perspectives of research are suggested to focus on a hybrid-based construction [25] which is crossover two different models in graph sequencing constructions. It will be supposed to improve the assembler's performance via combining advantages from both of motioned models. A hybrid graph sequencing constriction between OLC and greedy graph, that combines reads with different quality from different sequencers within the procedure named as "hybrid assembly", has been implemented nowadays. Particularly, the greedy overlap–based sequence assembler approach uses a greedy algorithm in the graph construction processing, which of the algorithm cannot be proved to output a globally optimal solution; however, such the approach is acceptable because the quality of the assembly meets or exceeds OLC assemblers via using a decent amount of computational hardware.

## 5. Conclusion

This paper addressed a new HMM-based method for DNA sequence assembly, which performs a high throughput sequencing rate with a relatively low cost. Besides these advantages, the proposed method focuses on introducing HMM into graph construction process in order to better assemble genome sequences, whose experimental results are proved that its performance meets or exceeds the state-of-art. Moreover, we point out that future hybrid method can involve new features in graph construction.

**References**

[1] Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE (2011) Landscape of next-generation sequencing technologies. Anal Chem 83: 4327–4341.
[2] Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55: 641–658.
[3] Z., Xia, et al. An audio digital watermarking algorithm in DCT domain for air-channel transmitting. Journal of University of Science and Technology of China, 2011(41), pp: 642 - 650.
[4] Helmy M, Sugiyama N, Tomita M, Ishihama Y (2012) Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics. Genes Cells 17: 633–644.
[5] D Chang, et al. Location based robust audio watermarking algorithm for social TV system. In Pacific-Rim Conference on Multimedia 2012 Dec 4 (pp. 726-738). Springer, Berlin, Heidelberg.
[6] X Zhang, etc. An Audio Steganography Algorithm Based on Air-Channel Transmitting. Journal of Wuhan University (Natural Science Edition), 2011, 57(6): 499 – 505.
[7] Helmy M, Tomita M, Ishihama Y (2011) Peptide identification by searching large-scale tandem mass spectra against large databases: bioinformatics methods in proteogenomics. Genes, Genomes and Genomics 6: 76–85.
[8] Zhou X, Ren L, Meng Q, Li Y, Yu Y, et al. (2010) The next-generation sequencing technology and application. Protein Cell 1: 520–536.
[9] Liu L, Li Y, Li S, Hu N, He Y, et al. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.
[10] Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, et al. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18: 810–820.
[11] X Zhang., and D Chang. The integrated scheme of High-Definition Environmental-Protection Multimedia Intelligent Conference System. IEEE 2nd International Conference on Mechanic Automation and Control Engineering (MACE). 2011, July: pp. 4576-4579.
[12] Chaisson M, et al. Fragment assembly with short reads. Bioinformatics, 2004(20): 2067–2074.
[13] Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Res, 2008(18): pp. 324–30.
[14] DiGuistini S, Liao NY, Platt D, Robertson G, et al. (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol 10: R94.
[15] Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30: 693–700.
[16] Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res 17: 1697–1706.
[17] Gonnella G, Kurtz S (2012) Readjoiner: a fast and memory efficient string graph-based sequence assembler. BMC Bioinformatics 13: 82.
[18] Li R, Zhu H, Ruan J, Qian W, Fang X, et al. (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20: 265–272.
[19] Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12: 671–682.
[20] Kao WC, Chan AH, Song YS (2011) ECHO: a reference-free short-read error correction algorithm. Genome Res 21: 1181–1192.
[21] Xia Z., etc. An audio digital watermarking algorithm transmitted via air channel in double DCT domain. IEEE International Conference in Multimedia Technology (ICMT), 2011: pp. 2926-30.
[22] Zhang, X., Chang, D., et al. Sonic Audio Watermarking Algorithm for Air-transmission. In Proceedings of the 2011 Third International Workshop on Education Technology and Computer Science-Volume 02 (pp. 110-113). IEEE Computer Society. 2011, March.
[23] Zhao, Y., et al. A robust audio sonic watermarking algorithm oriented air channel. IEEE International Conference on Computational and Information Sciences (ICCIS), 2011. pp: 53-57.
[24] D Chang, X Zhang, Y Wu. A Multi-Source Steganography for Stereo Audio. Journal of Wuhan University (Natural Science Edition). 2013;3: 277-284.
[25] Ding, Liang., et al. Efficient Learning of Optimal Markov Network Topology with k-Tree Modeling. arXiv preprint arXiv:1801.06900. 2018, Jan.