JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 7, 2022

Pp. 1–15
DOI: 10.1089/cmb.2022.29066.pp

Summary

Open camera or QR reader and scan code to access this article and other resources online.

Michael Waterman's Contributions to Computational Biology and Bioinformatics

PAVEL PEVZNER,[1] MARTIN VINGRON,[2] CHRISTIAN REIDYS,[3] FENGZHU SUN,[4] and SORIN ISTRAIL[5]

ABSTRACT

On the occasion of Dr. Michael Waterman's 80th birthday, we review his major contributions to the field of computational biology and bioinformatics including the famous Smith-Waterman algorithm for sequence alignment, the probability and statistics theory related to sequence alignment, algorithms for sequence assembly, the Lander-Waterman model for genome physical mapping, combinatorics and predictions of ribonucleic acid structures, word counting statistics in molecular sequences, alignment-free sequence comparison, and algorithms for haplotype block partition and tagSNP selection related to the International HapMap Project. His books *Introduction to Computational Biology: Maps, Sequences and Genomes* for graduate students and *Computational Genome Analysis: An Introduction* geared toward undergraduate students played key roles in computational biology and bioinformatics education. We also highlight his efforts of building the computational biology and bioinformatics community as the founding editor of the *Journal of Computational Biology* and a founding member of the International Conference on Research in Computational Molecular Biology (RECOMB).

Keywords: RNA structure and word counting, sequence alignment, sequencing assembly.

1. INTRODUCTION

Michael Waterman grew up on a ranch in southwestern Oregon. During his childhood, he cared for cattle and sheep and worked long hours on the ranch. He did not have many books to read and his

[1]Department of Computer Science and Engineering, University of California San Diego, San Diego, California, USA.
[2]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany.
[3]Department of Mathematics, Biocomplexity Institute & Initiative, University of Virginia, Charlottesville, Virginia, USA.
[4]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California, USA.
[5]Department of Computer Science, Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA.

grades in elementary school were less than satisfactory. The childhood experiences were documented in his 2016 memoir, *Getting Outside: A Far-Western Childhood* (Waterman, 2016).

His life was completely changed in 1960 when he enrolled in Oregon State University (OSU) where he majored in mathematics. He was a first-generation college student at the time and received his BS and MS degrees in mathematics in 1964 and 1966, respectively. During college he was fascinated by the breadth of knowledge and absorbed himself in reading and learning from many books ranging from philosophy to literature in addition to mathematics. He has established ''The Michael and Tracey Waterman Scholarship'' to support the studies of other first-generation college students at OSU.

He then went on to PhD studies in probability and statistics at Michigan State University and received his PhD degree in 1969. After graduation, he became a faculty member in the mathematics department at Idaho State University from 1969 to 1975. He then moved to Los Alamos National Laboratory in 1975 and became interested in the analysis of molecular sequences using mathematical and computational tools. In 1982 he moved to the University of Southern California (USC) as a professor of mathematics, computer science, and biological sciences. He retired from USC in 2019 and is an emeritus professor. He is currently a distinguished research professor at University of Virginia.

Professor Waterman is a widely acknowledged creative force whose work was pivotal in the establishment of the field of computational biology and bioinformatics. He was among the first innovators to apply complex mathematical and computational tools to the daunting challenges in DNA and protein sequence analyses. Among the many contributions he has made, the Smith–Waterman algorithm for local sequence alignment, the Lander–Waterman model for physical mapping, and the De Bruijn graph approach for sequence assembly were the most celebrated. However, Professor Waterman's work has had even further influences. He developed fundamental algorithms needed to analyze the complex secondary structure of ribonucleic acid (RNA) molecules; approaches to examine random-sequence matching scores and patterns; algorithms for declumping sequence alignments; algorithms for parametric alignment; algorithms for restriction and optical mapping, procedures for identifying haplotype blocks and tagSNPs; methods for alignment-free genome comparison; and numerous other important advances in the rapidly growing and evolving field of computational biology and bioinformatics.

In addition to research, Waterman has also advanced the field through his leadership in education, training, and the vigorous development of the computational biology community. In 1988 Waterman edited an influential book, *Mathematical Methods for DNA Sequences* (Waterman, 1989b) that included many important topics in molecular sequence analyses. In 1995 he published a seminal landmark book, *Introduction to Computational Biology: Maps, Sequences and Genomes* (Waterman, 1995). This textbook immediately led to a proliferation of university courses in this field. He also co-authored in 2007 a book, *Computational Genome Analysis: An Introduction* (Deonier et al., 2005), geared toward undergraduate students and others who did not yet have deep mathematical and statistical training. Advancing the field in another key respect was Professor Waterman's work as founding editor of *Journal of Computational Biology*, a dominant journal for mathematical and computational algorithms for genome analysis. In 1997, with Sorin Istrail and Pavel Pevzner, Professor Waterman co-founded RECOMB—Research in Computational Molecular Biology, a leading conference in the field that brings together annually the people and cultures of mathematics, statistics, computer science, and biological sciences. Twenty-five years after its initiation, RECOMB continues to be the dominant venue stimulating rapid innovation, advances, and expansion of the field.

In this review several of Waterman's former students, postdoctoral fellows, and colleagues highlight and review his major contributions to the field of computational biology.

2. THE ART OF PROBLEM FORMULATION (PAVEL PEVZNER)

When I teach sequence comparison, I first introduce the ''three-floor'' recurrence for the global sequence alignment between two strings. It describes how to compute the alignment score $s_{i,j}$ between the prefix of the first string of length i and the suffix of the second string of length j:

$$s_{i,j} = \max \begin{cases} s_{i-1,j} + \delta(v_i, -) \\ s_{i,j-1} + \delta(-, w_j) \\ s_{i-1,j-1} + \delta(v_i, w_j). \end{cases}$$

Afterward, I make a big deal about ''0'' in the Smith-Waterman ''four-floor'' recurrence (**?**) for the local sequence alignment:

$$s_{i,j} = \max \begin{cases} 0 \\ s_{i-1,j} + \delta(v_i, -) \\ s_{i,j-1} + \delta(-, w_j) \\ s_{i-1,j-1} + \delta(v_i, w_j). \end{cases}$$

My students do not seem to be impressed—for them, adding ''0'' is merely a decoration on the familiar recurrence for the global alignment—it accounts for the ''free ride'' to the starting point of the local alignment path in the dynamic programming graph:

At this point, I slow down and explain that, to truly appreciate the ''power of 0,'' they have to travel back to the 1980s and read some articles that struggled to solve a similar ''local'' comparison problem without much success. I happened to read these (rather complex) articles when I started working on my PhD in 1985 and, at that time, I had no clue about the magic ''0'' in the formula above. I started by learning the key idea behind the global alignment algorithm that was discovered and re-discovered many times—Sankoff (2000) provided an excellent review of the history of these discoveries in bioinformatics and beyond. Later, I learned that this key idea can be traced back to 1938 when Robinson (1938) described an algorithm for constructing the Young Tableau of a permutation that later led to the Robinson–Schensted–Knuth algorithm (Schensted, 1961; Knuth, 1970).

Although I quickly learned how to construct the global alignment, it was absolutely unclear how to solve (or even formulate) the local alignment problem (Sellers, 1980). Reading the local alignment article by Smith and Waterman (1981b) was a revelation—how it can be so simple and how possibly nobody saw these free rides before!

I tell my students that one of the most difficult challenges in bioinformatics is the art of problem formulation (and re-formulation!). It is the new elegant formulation of the local alignment problem (and a switch from minimizing the edit distance to maximizing the sequence similarity) that was the most difficult challenge. Only after the problem was formulated, it became clear that ''free rides'' dictate adding an extra floor to the original three-floor recurrence.

Michael Waterman perfected the art of problem formulations in bioinformatics and pioneered unexpected avenues for their solutions. Genome assembly is another area where Waterman came up with a surprising algorithmic approach that some colleagues thought had no chance to succeed: in fact, the now famous Idury and Waterman genome assembly article (Idury and Waterman, 1995) had very few citations in the first decade after its publication.

Biologists still cannot read genomes from the beginning to the end (like we read a book)—instead, they generate short fragments of the genome called *reads* and try to assemble them into a complete genome. Genome assembly is the largest (although, one-dimensional) jigsaw puzzle humans ever faced.

Children usually assemble puzzles by trying all possible pairs of pieces and putting together pieces that match. Before Idury and Waterman published their article in 1995 (Idury and Waterman, 1995), biologists assembled genomes in a similar way: all previous assembly efforts were based on the *overlap–layout–consensus* approach that represents each read as a vertex in the *overlap graph*. Vertices v and w in this graph are connected by an edge if a sufficiently long suffix of v coincides with a prefix of w (or, to account for errors in reads, is very similar to a suffix of w). The genome assembly problem is reduced to finding a *Hamiltonian path* in the overlap graph, that is, a path that visits each vertex (read) once.

To address the limitations of the overlap–layout–consensus approach, Idury and Waterman tried to transform genome assembly from a difficult Hamiltonian path problem in the overlap graph to an easy *Eulerian path* problem in a different graph. To construct this different graph from a set of DNA reads, they abandoned the traditional puzzle assembly approach by deciding to never match the pairs of fragments! You may be wondering how one can possibly assemble a puzzle without doing it.

Idury and Waterman (1995) drew inspiration from *Sequencing By Hybridization*, an alternative DNA sequencing technology aimed at generating all short k-mers from a genome and further assembling them

(Drmanac et al., 1989). Instead of the overlap graph, one can construct the *de Bruijn graph* from all k-mers where vertices represent all $(k-1)$ -mers and each k-mer corresponds to an edge connecting its prefix $(k-1)$-mer with its suffix $(k-1)$-mer. An Eulerian path in the resulting graph spells out the unknown genome and thus solves the genome assembly problem (Pevzner, 1989).

However, the described approach requires knowledge of all k-mers from the genome. A typical state-of-the-art assembly project in 1995 would generate ≈ 2000 reads of length ≈ 300 nucleotides for a ''genome'' (clone) of length $\approx 45{,}000$ nucleotides—less than 5% of 300-mers needed for constructing the de Bruijn graph on all 300-mers from this genome. To recover all k-mers from a genome and thus enable the de Bruijn graph construction, Idury and Waterman proposed a counterintuitive approach—breaking each DNA read into shorter overlapping k-mers (e.g., 50-mers) so that all or nearly all k-mers from the genome appear as one of the broken pieces. Although this ''*assembling a puzzle by breaking it into smaller pieces*'' approach indeed enables the construction of the de Bruijn graph, read breaking results in information loss and, at the first glance, appears to be a desperate and ill-conceived attempt at hammering the wrong nail (applying the de Bruijn graph to genome assembly). Not to mention that the de Bruijn graph approach was designed for error-free k-mers while DNA reads have errors.

Although most people were skeptical about the de Bruijn graph approach to genome assembly back in 1995, 6 years later, Pevzner et al. (2001) showed how to error-correct reads making them nearly error-free (and thus perfectly suited for the de Bruin graph approach) and how to restore information lost during breaking reads into shorter k-mers. By 2010, nearly all genome assemblers followed the de Bruijn approach—the vast majority of genomes known today have been assembled using the approach inspired by the Idury and Waterman (1995) article.

3. SEQUENCE ALIGNMENT ALGORITHMS AND STATISTICS (MARTIN VINGRON)

3.1. Sequence alignment algorithms

The earlier section already introduced the ''three-floor'' and ''four-floor'' recurrence rules for global and local alignment, respectively. Indeed, the ''four-floor'' Smith–Waterman algorithm with the famous zero, published in 1981, has become a ubiquitous tool not only in computational biology, but in molecular biology in general. It is less known, however, that in 1981, there appeared two ''Smith-Waterman'' articles: The one on local alignment describing the famous algorithm (Smith and Waterman, 1981b) and another one on the equivalence of similarity and distance in sequence alignment (Smith and Waterman, 1981a). Although the second one has not become as famous as the first one, both articles solve very fundamental problems.

Sequence alignment algorithms are beautiful not only for the application of dynamic programming, but maybe more so by virtue of the evolutionary model behind sequence alignment. When initially the question was to define a distance on sequences with the idea of quantifying the amount of change to convert one sequence into the other, evolution provides us with the mechanism why mutations, insertions, and deletions are the right operations to look at. Thus, a global alignment between two sequences provides not only a minimal distance in terms of these operations, it actually suggests that these operations, in some order, might have happened in history. On the contrary, one can interpret sequence alignment as maximizing a scoring function, be it the sum of matches or the sum the physical similarity values between amino acids (always reduced by the gap penalties).

During the initial years of the development of alignment algorithms, it was an open question which of these viewpoints, minimization of a distance or maximization of similarity, was preferable. In ''Comparison of Biosequences'' [Smith and Waterman (1981a); see also Smith et al. (1981)] Temple Smith and Michael Waterman showed that under very general assumptions, for a distance metric there also exists a similarity scoring scheme that will yield the same optimal alignment. The proof relies on the ''alignment invariant,'' which says that however one aligns two sequences, the sum of number of indels plus twice the number of matches will be equal to the sum of lengths of the two sequences. For amino acids the formulation is more involved, but in essence similar. This settled a whole line of discussion and provided a unified basis for the field.

As pointed out in the earlier section, for many years the community has searched for a good algorithm to find local alignments, often times basing the search on minimizing evolutionary distance. In the local alignment case the alignment invariant does not hold and it was unclear from which side to attack the problem. Remember that the recursion for global alignment considers the possibilities that an alignment

gets extended with another matching pair, or that an indel is added to either of the sequences. This did not suffice to locate a common segment between two sequences that before and/or after this common segment do not share any more similarity. The first half of the solution was the zero added to the choices, as also described in Pavel Pevzner's contribution to this article. This zero takes care of skipping from the beginning of the sequences to the common segment. It gets combined with starting the backtracking from the highest point of the matrix, which disconnects the common segment from the ends of the sequences.

The less prominent ingenious trick here is that in this case distance versus similarity makes all the difference. The Smith–Waterman local alignment algorithm works because it *maximizes* a segment score. It is like seeing a chain of mountains rise out of the morning fog: The zero defines the fog line above which the mountains become visible. The alignment algorithm, like a hiker, finds the top of the mountain. There the optimal path ends and from there the backtracking, the hiker's descent, starts.

Local alignment has remained in the focus of Mike Waterman's research. A very prominent type of problem still could not be solved. Eukaryotic genes come in pieces, exons, which tend to be more conserved than the regions in between the introns. Thus when comparing the genomic sequences of two homologous gene regions, the one local alignment is usually given by the best matching pair of exons and the user will remain blind to the matches among the remaining exons. Those matches would constitute suboptimal solutions and in general they would be clearly visible in a dot plot. But then again, a suboptimal alignment in a strict sense might be one that is distinguished from the optimal one only by some uninformative wiggle. In a 1987 article in *Journal of Molecular Biology* (Waterman and Eggert, 1987), Michael Waterman showed how to erase from the edit matrix all traces of an alignment such that the search for a next best one could be continued without reiterating what one had computed before. This mechanism allows one to start with the best local alignment, erase it together with all its wiggles, collect the next best alignment, and so on. At the same time this ingenious algorithm provided the basis for the statistical treatment of local alignment that will be discussed hereunder.

Several other fundamental contributions shall be mentioned only shortly. The question of the dependence of the optimal global alignment on the parameters (one or two gap penalties) under which it was calculated is at the core of parametric alignment. The key observation here is that because the scoring function depends linearly on the parameters, one alignment is optimal within a convex parameter region. In Waterman et al. (1992), Waterman presented an algorithm to compute the resulting tessellation under variation of the parameters. Again, this discrete algorithm developed a probabilistic side later, which will be discussed hereunder. Waterman also contributed to the multiple alignment problems (Waterman and Perlwitz, 1984) and to pattern finding in biological sequences (Waterman et al., 1984; Galas et al., 1985).

3.2. Statistical significance of sequence alignment score

Alignment algorithms search for the optimal alignment between two sequences, that is, the one maximizing a score that is additively composed of the match qualities and the (negative) gap penalties. This makes it nontrivial to assign a statistical significance to the optimal alignment score, because this value tends to be large already owing to the maximization. As such, the naive approach of describing any random variable as being normally distributed will fail. This was reflected in the rule of thumb propagated in the 1980s, that an alignment score needs to be at least 4, better 6 standard deviations above the mean score of random sequence pairs. Although such a perspective on the problem stems from the normal assumption, were the random variable really normally distributed this would happen very rarely and the recommendation reflects the fact that the scores tend to become much larger than conceivable under a normal distribution.

Together with developing algorithms for sequence alignment, Waterman realized early that one needs to understand the distribution of the random variable ''alignment score.'' In a seminal 1985 article in *Nucleic Acid Research*, Waterman together with Smith and Burks (Smith et al., 1985) described empirically the distribution of scores one obtains upon searching a database of sequences. There the authors pointed out that runs of matches between two sequences, allowing for shifting the sequences with respect to each other, generalize head runs in coin tossing, which had been studied by Erdös and Rényi. Building on this relationship, Waterman, mostly together with Gordon and Arratia, published a series of articles where they showed that this viewpoint leads to a description of the extreme value behavior of what at the time was called ''segment score.''

Many problems had to be solved along the way. In 1985, Arratia and Waterman showed how to generalize the Erdös-Rényi analogy to shifts and thus to sequence comparison (Arratia and Waterman,

1985). Furthermore, unlike head runs, which are by definition uninterrupted, matching segments between sequences may contain mismatches. This motivated a study of interrupted head runs in ''An extreme value theory for sequence matching'' published in 1986 (Gordon et al., 1986). In a 1990 article in *The Annals of Statistics*, this was all brought together under the title ''The Erdos-Rényi law in distribution, for coin tossing and sequence matching'' (Arratia et al., 1990b).

With this, they had a handle on unraveling the distributional law behind local sequence alignment. However, the indels were still a problem. Together with Richard Arratia, Michael Waterman discovered that the gap penalties introduce a phase transition in the behavior of alignment score (Arratia and Waterman, 1994): when computing alignment scores for random sequence pairs of increasing length, under weak gap penalties these score will grow linearly, whereas when heavily penalizing gaps, alignment score will grow logarithmically with the lengths of the sequences compared. In addition—and this is the phase transition—there exists a boundary at which the one regime changes to the other. In an article in *PNAS*, together with Gordon and Arratia (Waterman et al., 1987), a figure shows two typical alignments: One in the permissive gap regime where the alignment is all torn apart and stretches over the entire length of the two sequences, and a second one where the aligned region is compact with only occasional indels, except for the long gaps leading to this matched region.

The extreme case of a compact alignment is a segment alignment without any gap. The behavior of such a segment alignment score is well described by an extreme value distribution. Together with the phase transition, this suggests that the extreme value behavior should also hold for any compact, local alignment given that the penalties are sufficiently high so as to remain on the logarithmic side of the phase transition. In Waterman and Vingron (1994) and Vingron and Waterman (1994) we showed empirically that this holds true. In addition, there is a connection to the declumped suboptimal alignments computed by the Waterman–Eggert algorithm. The number of suboptimal alignments above a particular threshold yields the intensity of a Poisson process, which in turn lets us estimate the statistical significance of the optimal alignment. This follows the theory laid out in Goldstein and Waterman (1992), where the authors integrated earlier work on Poisson approximation and the Chen–Stein method (Arratia et al., 1990a).

4. COMBINATORICS AND PREDICTION OF RNA STRUCTURES (CHRISTIAN REIDYS)

4.1. Prediction of RNA secondary structures

Almost four decades ago Michael Waterman pioneered the combinatorics and prediction of the then rather exotic RNA secondary structures. On the one hand, an RNA molecule is described by its primary sequence, a linear string composed of the nucleotides A, G, U, and C. On the other hand, RNA, being less structurally constrained than its chemical relative DNA, does fold into tertiary structures.

The notion of RNA secondary structures was proposed in 1960 (Fresco et al., 1960). The central question was centered around the prediction of the minimum free energy (MFE) secondary structure. One prominent early prediction method was owing to Tinoco et al. (1971) and based on the base pairing matrix. This method was widely used and formed the backbone of many subsequent algorithms. However, algorithms following this paradigm require an exhaustive search over all combinations of helixes and consequently are infeasible for long RNA sequences.

In Waterman (1978), Waterman formalized the mathematical concept of RNA secondary structures. He provided a graph theoretical definition and proved various enumeration results, loop classification, and decomposition. These concepts shaped the field and played a fundamental role in countless contributions on RNA and protein secondary structures. In addition he introduced a classification of secondary structures as a basis for a new and efficient MFE folding algorithm. Waterman then introduced key concepts that classify substructures through loops, bulges, interior loops, joins, ladders, and tails. These definitions constitute now ''classical'' terms. To measure the algorithmic complexity of a secondary structure, he introduced the notion of the order of a secondary structure. The order of a secondary structure can be viewed as the ''depth'' of the structure with respect to recursive resolving of hairpins. On top of this he conducted an analysis on energy functions, producing very general results, and laying the foundation for loop-based energy models. Current nearest neighbor energy models can be viewed as a simplification of the general model introduced in this article. It is fair to say that this contribution had profound conceptual influence on mathematics and algorithms for secondary structure prediction. It not only lays the

foundation of mathematical studies on RNA secondary structures but also shapes subsequent work on folding algorithms.

The same year a joint work with Temple Smith was published (Waterman and Smith, 1978). Key achievements therein were a rigorous mathematical analysis on the prediction of RNA secondary structure through energy optimization: an efficient dynamic programming routine that allowed a search over the entire configuration space of the RNA molecule not possible by earlier methods and the direct inclusion of the nearest neighbor or stacking energies. The article made clear that the free energy of a structure can be viewed abstractly as a distance score. Energy functions for each type of substructure were given and the total free energy was computed as the sum of the free energy associated with these substructures. The compatibility of energy and combinatorial recursion allowed to compute the MFE. Finally, a back tracking routine, specifying a particular secondary structure that achieved this MFE was presented. This result later constituted the core of John McCaskill's partition function contribution (McCaskill, 1990).

Waterman and Smith (1986) provided the first polynomial time dynamic programming algorithm to predict general MFE structures for a given sequence. The framework established here formed the intellectual core of a plethora of RNA folding algorithms. They presented here fundamental ideas reducing the theoretical and practical computational complexity of dynamic programming algorithms.

Finding the consensus structure of multiple homologous RNA sequences is crucial to the understanding of their functions. It is challenging for MFE methods to handle many sequences simultaneously. Waterman (1989a) used comparative analysis to find the consensus structure of multiple homologous RNA sequences. This article laid the foundation for a vast number of articles. This framework recruits exclusively comparative analysis and requires no additional knowledge about the thermal dynamic stability of the structures. The key idea behind this method is that important features, such as specific bases or helices, have been conserved over the course of RNA evolution. It exhibits two main components: alignment across sequences to detect conserved bases and alignment within a sequence to detect helices.

4.2. Combinatorics of RNA structures

Stein and Waterman (1979) generalized the quadratic recursion of the classic Catalan numbers and Motzkin numbers, the first time to connect them with RNA secondary structure. The concepts were motivated by enumerating secondary structures with specific biological constraints, related to loop energies, such as minimum arc length. In a sense, this article (Stein and Waterman, 1979) connected combinatorics to molecular biology and marks the advent of some version of discrete biomathematics.

Waterman (1979) conducted an analysis on two types of secondary structures: hairpins and cloverleaves. He presented the asymptotics of the number of generalized cloverleaves, namely, the ''loop'' region containing three or more hairpins.

Howel et al. (1980) provided various algorithms and techniques for computing generating functions of RNA configurations. They established in particular the connection between the computation of generating functions with formal grammar as the basis for all stochastic folding algorithms. A key result of the article was the computation of the bivariate generating function of secondary structures that are compatible with a given sequence. Penner and Waterman (1993) presented seminal work providing a topological framework for the space of RNA secondary structures. The main result is about the sphericity of the topological spaces of both arbitrary secondary structures and binary secondary structures.

The contribution by Schmitt and Waterman (1994) is of exceptional beauty. Its subject is the enumeration of the number of RNA secondary structures of a given length with a fixed number of base pairs. This is performed by establishing a one-to-one correspondence between secondary structures and linear trees. A linear tree is a rooted tree together with a linear ordering (left to right) on the set of children of each vertex in the tree. A duality operator on trees is presented, which explains a symmetry in the numbers counting secondary structures. In the context of connecting combinatorics and molecular biology, the article (Schmitt and Waterman, 1994) gave rise to a variety of secondary structure comparison methods with important implications to the structural evolution of RNA.

RNA pseudoknot structure plays important roles in various molecular functions and structural prediction involving pseudoknots remains a challenge. One particular question here is how to conceptualize the notion of pseudoknot RNA. In Andersen et al. (2013), Waterman provided a path to such a conceptualization through fatgraphs. They established the classification of RNA pseudoknot structures via topological genus and link RNA enumeration problems with the geometry of Riemann's moduli space. Using enumeration

results of Harer–Zagier on chord diagrams, the authors computed the generating function of structures of fixed genus and minimum stack size with a given number of nucleotides in which no two consecutive sites are paired. The topological classification turns out to have connections with algorithmic complexity. In particular, the recursion presented in Andersen et al. (2013) is the basis of an efficient algorithm that involves structures with bounded topological complexity (Huang and Reidys, 2016; Barrett et al., 2019).

5. THE LANDER-WATERMAN MODEL FOR PHYSICAL MAPPING AND OTHER TOPICS (FENGZHU SUN)

5.1. Modeling the progress of physical mapping strategies

Before the dawn of genome projects of various organisms in the early 1980s, there were many debates about the optimal ways to sequence the genomes. Physical mapping strategies used sequence fragments, genomic markers, or restriction sites to map the relative positions of nucleotide bases or genomic fragments along the genomes. Common problems for physical mapping include the following: (1) how to link different sequence fragments into contigs composed of overlapping fragments? (2) how do the fraction of genomes belonging to characterized fragments, the number of contigs, contig lengths, and so on, change with the sequence coverage defined as the average number of times a genomic position was characterized? and (3) what sequence coverage should be used in genome projects? These are essential questions before carrying out any physical mapping experiments.

To answer these fundamental questions, Lander and Waterman (1988) developed the first mathematical model for physical mapping by fingerprinting random clones. The fingerprints of clones were characterized by their restriction fragment lengths at the time of study and two clones were declared as overlapping if they shared some overlapping fingerprints. In the model, the starting positions of the clones were modeled by homogeneous/inhomogeneous Poisson processes with clone lengths following a certain distribution. Two clones were declared overlapping if they shared a fraction of the clone lengths. Based on this model, Lander and Waterman (1988) derived explicit formulas for the fraction of genomes that belonged to characterized clones, the distributions of the numbers of contigs, contig lengths, and gap lengths. Based on the results from this relatively simple model, they suggested that a coverage of six was appropriate for most genome projects to balance cost–benefit ratio. The results from the study laid important foundations for most of the genome projects.

Another approach to link overlap clones was through anchoring using unique markers where overlap was determined by DNA sequences. Two clones were declared overlapping if they shared some common anchors. Arratia et al. (1991) developed a mathematical model for the progress of physical mapping by anchoring random clones. In addition to the assumptions given in the Lander–Waterman model (Lander and Waterman, 1988), the authors further assumed that the anchor markers followed another Poisson process. Paired-end sequencing was proposed in the early 1990s and Port et al. (1995) investigated the mapping progress based on the Lander–Waterman model using paired-end sequencing.

The Lander–Waterman model for physical mapping has since been applied to many different problems including contig binning in microbial communities and evaluation of sequencing libraries to name just two applications.

5.2. Word count statistics and alignment-free sequence comparison

Stochastic modeling of molecular sequences has a long history and has been widely used to characterize and compare molecular sequences. Investigations on the distributions of the number of occurrences of word patterns or k-mers, consecutive sequences of k letters, play important roles in molecular sequence analysis. There are two ways of counting k-mers: overlapping and nonoverlapping counts. In overlapping word counts, the k-mers in a sequence are counted consecutively from the beginning to the end. In nonoverlapping word counts, once a k-mer is observed in a sequence, one goes to the end of the k-mer and starts the counting process again.

Molecular sequence comparison has been one of the key problems in computational biology. The primary approach for sequence comparison is through alignment. However, alignment has limitations for some problems such as the comparison of regulatory regions and the comparison of genomes/metagenomes sequenced by next-generation sequencing. Regulatory regions are not highly conserved and thus it is

challenging to align gene regulatory regions. Alignment-free sequence comparison methods based on k-mer frequencies of the sequences of interest have the potential to be used for the comparison of such sequences. Torney et al. (1990) proposed to use the number of shared k-mers between two sequences referred as D_2 to measure the similarity between two sequences. D_2 can be calculated by the sum of products of $X_{\mathbf{w}}$ and $Y_{\mathbf{w}}$ over all the k-mers, where $X_{\mathbf{w}}$ and $Y_{\mathbf{w}}$ are the numbers of occurrences of k-mer $\mathbf{w}$ in the first and second sequences, respectively. Lippert et al. (2002) investigated the distribution of D_2 under a variety of different scenarios and showed that it is dominated by the variance of the number of occurrences of each k-mer in individual sequences.

Through the theoretical studies, it was hypothesized that the power of D_2 for detecting the relationship between sequences could potentially be increased by replacing $X_{\mathbf{w}}$ and $Y_{\mathbf{w}}$ with $X_{\mathbf{w}} - E(X_{\mathbf{w}})$ and $Y_{\mathbf{w}} - E(Y_{\mathbf{w}})$, respectively, under certain probability models for the sequences. This hypothesis was proved theoretically and by simulations in Reinert et al. (2009), Wan et al. (2010), and Liu et al. (2011). For applications to the comparison of genomic sequences, they standardized the resulting statistics to d_2^* and d_2^S based on different ideas of normalizing D_2. The studies on alignment-free genome and metagenome comparisons based on d_2^* and d_2^S were reviewed in Song et al. (2014) and Ren et al. (2018).

5.3. Sequencing accuracy, haplotype block partition, and tagSNP selection

Sequencing technologies played key roles in most genomic studies and it is essential to estimate the sequence quality from these technologies. Churchill and Waterman (1992) developed a mathematical model for sequencing and an expectation-maximization algorithm to estimate the error rates during sequencing and to give the posterior probability of bases in the consensus sequence. Kim et al. (2007) extended the model in Churchill and Waterman (1992) to diploid genomes and Li et al. (2004) developed a computational method to infer haplotypes from sequenced fragments based on the extended model.

In the early 2000s, several research groups observed long range linkage disequilibrium (LD) among single nucleotide polymorphisms (SNPs) in many human populations raising the possibility of using LD for genomewide association studies of complex diseases. At the time there were extensive interests in dividing the human genome into blocks such that SNPs within each block are in high LD and SNPs between different blocks have limited LD. In each block a set of SNPs referred as tagSNPs are selected such that the tagSNPs can explain at least a given fraction of haplotype diversity in the block. The diversity in each block can be defined in various ways depending on the researchers' interest. Zhang et al. (2002b) rigorously formulated the problem as minimizing the total number of tagSNPs and developed a dynamic programming algorithm to solve this problem for haplotype data. By integrating haplotype inference based on genotype data from Jun Liu's group with the dynamic program algorithm for haplotype block partition and tagSNP selection, the dynamic programming algorithm was further extended to deal with both haplotype and genotype data (Zhang et al., 2004) and with limited resources (Zhang et al., 2003).

They further showed that the power of association studies using tagSNPs was only slightly lower than that using all SNPs (Zhang et al., 2002a). The series of algorithms for haplotype block partition and tagSNP selection were integrated into a user-friendly software package, HapBlock, which was used widely by the scientific community (Zhang et al., 2005). The research on haplotype block partition and tagSNP selection had significant impacts on the International HapMap project and laid the foundation for the Center for Excellence in Genomic Sciences at USC from 2003 to 2015.

6. MICHAEL WATERMAN: A NATIONAL TREASURE! (SORIN ISTRAIL)

I write with humility, for how can I do justice in discussing the achievements of the father of the field, whose pioneering advances in applied mathematics analysis of DNA, intertwining continuous and discrete mathematics, became quintessential for computational biology. To speak of Professor Waterman is to speak of what matters most in this new and exciting field. He stands on high ground with an unmatched body of research, fundamental for every aspect of the discipline, evolutionarily intertwined with the demi-god Manhattan Project mathematicians who were in love with biology.

Santa Fe, the Manhattan Project, and John von Neumann intertwine in New Mexico, and we can trace some evolutionary roots of computational biology from there. An evolutionary trajectory of theories, collaborations, and mentoring in this ''land of enchantment'' reaches to one common ancestor node: John

von Neumann, who proposed a research program aimed at a new computation and information theory essential for modeling the biological systems of the cell. The overarching goal of his new theory was the unification of continuous and discrete mathematics through the concept of statistical thermodynamic error. The two pillars to be unified, using statistical thermodynamics, were continuous mathematics, namely mathematical analysis, as he put it, ''the technically most successful and best elaborated part of mathematics,'' and discrete mathematics, which ''deals with rigid, all-or-none concepts, and has very little contact with the continuous concept of the real or complex number … technically most refractory parts of mathematics … by the nature of its approach, cut off from the best cultivated portions of mathematics, and forced onto the most difficult part of mathematical terrain, into combinatorics.'' Professor Waterman's work—with its pioneering influence, trailblazing new areas of research through a wide range of computational and mathematical methods applied to biological problems, weaving statistical and algorithmic methods toward a genuine statistics-powered continuous-discrete hybrid—is a singular body of work closest, as any other, to von Neumann's herculean vision and research program.

I have known Mike for 30 years. Our collaboration began in 1992, when, two weeks after joining Sandia National Laboratories in Albuquerque, New Mexico, I participated in a Rutgers University tutorials Workshop designed to attract computer scientists and mathematicians to work on the daunting research problems of the Human Genome Project—a project started by the visionary Charles DeLisi at the U.S. Department of Energy (DOE). Mike was a Workshop lecturer. At Sandia, I was charged with establishing the Computational Biology Project, and I had been directed to work toward attracting Mike to Sandia by Dr. Fred Howes, the funding director of my new project and director of the Applied Mathematics Program (founded at DOE by John von Neumann, who was able to secure a 2000-fold increase of applied mathematics funding at the National Labs). We succeeded, and Mike's ensuing visits were instrumental in establishing our three-decade-long collaboration. With Professor Pavel Pevzner of University of California, San Diego, we three collaborated as the leaders of a variety of projects providing computational and mathematical sciences foundations for the computational biology and bioinformatics community: the RECOMB conference, the *Journal of Computational Biology*, the MIT Press Computational Molecular Biology book series, and the Springer-Verlag Lecture Notes in Bioinformatics book series.

I remember vividly when in 1997 Mike took Pavel and me to the late Stan Ulam's house in Santa Fe, New Mexico, to bring flowers to his wife, Francois Ulam. It was an emotional visit, for it was as though we were making a connection to great mathematicians who preceded us. We were in Santa Fe to establish what has come to be known as the RECOMB conference—the Annual International Conference on RECOMB. It just so happened that von Neumann, involved in the war effort at Los Alamos, brought Stan Ulam to New Mexico; Ulam brought Bill Beyer to New Mexico; Beyer brought Mike to New Mexico, and, thanks to Mike, we were in New Mexico for the first RECOMB conference. There, Mike passed the torch from mathematicians in love with biology at the end of the 20th century and the start of the 21st. Flanked by Bill Beyer and Nick Metropolis—two members of the greatest generation of mathematicians in residence in New Mexico—Mike declared: ''This volume [the first RECOMB Proceedings volume] is proof positive of the vitality of a new discipline: Computational Biology.'' The founding Steering Committee of the conference included, in addition to we three founders, Richard Karp (UC Berkeley), Tom Lengauer (Max Plank), and Ron Shamir (Tel-Aviv). It was Ron who coined the name RECOMB.

This year, 2022, the conference will celebrate in San Diego its 26th year, and it truly has had an international journey: Santa Fe (1997); New York City; Lyon; Tokyo; Montreal; Washington, DC; Berlin; San Diego, CA; Cambridge, MA; Venice; San Francisco; Singapore; Tucson; Lisbon; Vancouver; Barcelona; Shanghai; Pittsburgh; Warsaw; Los Angeles; Hong Kong; Paris; Washington, DC (2019); followed by two pandemic years of virtual conferences in 2020 and 2021, and hopefully in person this year. RECOMB is without question one of the top two international conferences in computational biology and bioinformatics, the other being the Intelligent Systems for Molecular Biology Conference (ISMB). RECOMB owes a debt of gratitude to Mike, who back in 1997 was our flagship.

RECOMB, established as an Association for Computing Machinery (ACM) conference, successfully borrowed for this exciting domain of computational biology the most effective traits of the exceedingly competitive top computer science conferences. RECOMB selects annually an internationally recognized program committee composed of most active and distinguished researchers of the field and selects the top paper submissions through a rigorous refereeing and intense program committee teamwork. This effort, spanning many weeks, culminates in the selection of a very small fraction from the large number of high-quality submissions. RECOMB is recognized for meeting a set of scientific standards that, for such a highly

interdisciplinary field, require tremendous innovation and contributions. We have Mike to thank for this. He played a tremendous leadership role in all scientific aspects of the conference and has been a major contributor in designing the rules of scientific analysis that shape the advancement of the field by setting uncompromising standards; for selecting invited speakers of distinction—with special attention to recognizing local pioneers in the host country—and ensuring the fairness that is necessary in the confidential program committee work. Without question, Mike's contributions shaped the course of computational biology and bioinformatics toward the maturity of the field—a maturation that is reflected in RECOMB articles demonstrating significant innovation in computational and mathematical foundations of the field.

The conference also has been providing a major educational component for graduate students, postdoctoral students, and young faculty. The conference organizers are always making a major effort to secure funding for graduate students and postdoctoral students to attend in large numbers. The conference truly unified the field of computational biology and bioinformatics internationally, establishing RECOMB as the most prestigious level of achievement annually. The 35 RECOMB articles accepted annually (selected from 200 to 250 submissions) are hard-core computational biology's top 35 articles of the field.

The *Journal of Computational Biology* is without question the top journal of the area devoted to innovation in computational and mathematical sciences. As the founding editor, Mike's leadership was instrumental in reaching this milestone. In addition, Mike is a best-selling author of textbooks on computational biology. This reflects another shining dimension of his pedagogical philosophy and passion that resonated so well with generations of professors of this new and exciting field: he is our professor-in-chief. His students hold positions of highest stature in computational biology, bioinformatics, and biotechnology in several countries—United States, Germany, United Kingdom, France, and China, to name a few. In this respect, Mike has the most outstanding record of all professors of the field.

Mike, together with Temple Smith, invented at Los Alamos National Laboratory the most beautiful algorithm of computational biology, the Smith–Waterman local alignment algorithm. I teach it in my Algorithmic Foundations of Computational Biology course at Brown, and it takes me about six lectures to cover the topic, although, as simplicity is indeed the ultimate sophistication, it is a 4-line algorithm. It is one of the deepest topics in computational biology. The intertwining of statistics (e.g., random walks with negative drift, and parameters settings and phase transition) and algorithms (e.g., why an exact globally optimal maximization algorithm stops at the optimal local alignment [of logarithmic sequence size], instead of reaching the global alignment [linear sequence size]?) is truly von Neumannesque in harnessing the rigorous modeling of the fascinating but mysterious structure of the DNA sequence. I believe that the most cited article in science is the BLAST algorithm article, with about 75,000 citations to date. Though Mike is not an author of that article, the Smith–Waterman algorithm is at its heart. The ultimate tribute of an article is that it is so entrenched in the scientific culture that people do not cite it anymore. By now, the Smith–Waterman part of BLAST has been in use every second of every day by thousands of users for 30+ years, and in every genome assembly of every organism sequenced and assembled to date; the human genome employs billions of runs of the algorithm. All in all, the Smith–Waterman algorithm, with algorithmic speedups and deep statistical theory, define ''the practical'' for computational biology methods and genomics analysis.

Years ago, in 2003, in his office at USC, Mike gave me a book as a gift. ''To Sorin: You can know a man by his heroes. Best of everything Michael Waterman 17 December 2003,'' he wrote as a dedication. The book was the *Bulletin of the American Mathematical Society* volume 64, number 3, May 1958: John von Neuman, in memoriam. The volume's first page states: ''We do not erect statues of great scientists. Instead, the American Mathematical Society publishes this volume as a memorial to John von Neumann. Some of his friends describe his brilliant mind, his warm personality, his work which will live on in mathematics and in other sciences to which he has contributed so much.'' In the volume, the first article, by Stan Ulam, presents von Neumann's life and work from a background of personal acquaintance and friendship extending over a period of 25 years. In Tokyo, at RECOMB 2000, Mike introduced me to Eric Davidson. Over the next 15 years, till Eric passed in 2015, he became my mentor, my collaborator, and one of my closest friends. Eric was a tough scientist, critical and direct, but he had a subtle way of communicating important ideas through stories about his heroes.

In one such story (Istrail, 2016) at dinner, he told how his next-door colleague at Caltech, Max Delbruck, enticed him to study more mathematics. Thereafter, one of Delbruck's postdocs taught Eric the mathematics of Poisson statistical theory. I got the story's message, and took the Davidson-Delbruck challenge. Thereafter, I started to learn some wet-lab molecular biology procedures taught by Eric's postdocs. The

other message, about Poisson statistics, was crystal clear as well. I learned Poisson statistics from Mike's seminal articles: the Idury-Waterman (Idury and Waterman, 1995) genome assembly algorithm article and the Waterman et al. (Lander and Waterman, 1988; Arratia et al., 1991) statistical framework for genome assembly articles. Mike co-authored these gems of the technical discipline, and now I love teaching Poisson statistics in my classes. I love the Delbruck-Davidson and von Neumann-Ulam-Waterman branching processes and their intersections. Eric taught me a lot about molecular biology. In turn, I tried to teach him computer science, not always with much success. When we wrote our article, ''The regulatory genome and the computer'' (Istrail et al., 2007), as the 50 years homage to von Neumann's last book, ''The Computer and the Brain'' (1958), I taught Eric some Boolean logic and the principles of the von Neumann's architecture, present in every electronic digital computer then, and today. The two branching processes crossed: Eric devoured von Neumann.

I co-authored only one article with Mike (Istrail et al., 2004), and it is extraordinarily dear to me. It was a very global alignment algorithm article—this time, whole-genome-assembly-to-whole-genome-assembly alignment! I call it the ''lighthouse'' article. I was then Senior Director of Informatics Research at Celera Genomics/Applied Biosystems, and Michael Hunkapiller, the president of the company, together with Craig Venter, my former boss and at that time president of what would become the Craig Venter Institute, asked me to lead the effort to build the tools for the most comprehensive comparison of all human genome assemblies to date, both of the Human Genome Project public effort and of Celera. It was stressed to me that the article describing this analysis needed to be most comprehensive and accurate, leaving no room for shifting conclusions and hopefully containing beams of success, like a lighthouse. Mike became a collaborator of ours at Celera and, together with a large group of colleagues, we published the lighthouse article in the *Proceedings of the National Academy of Sciences* in 2004. As required by the publication, we worked with the NCBI leadership, and we transferred all Celera Genomics human assemblies to NCBI. Also, our genomics software tools, including the genome-to-genome alignment tool ATAC software package (Assembly-to-Assembly-Comparator) was placed in the public domain as open-source software in sourceforge.* Indeed, the results of our lighthouse article have stood the test of time.

New Mexico is an enchanted land of algorithms and flowers, a haven for romantic liaisons between science and the arts where each can reach unprecedented peaks. It is here Mike found inspiration for writing literature. Los Alamos' Nick Metropolis, whose algorithm is the most used/cited in the history of science and engineering, and Georgia O'Keeffe, whose paintings of desert flowers offer singular beauty, are characters in Mike's piece, ''Nick the Greek.'' Within his prose you hear the rivers of his native Oregon, see the brightness of the light, and feel the fresh high-desert air of New Mexico that he calls skiing the sun.

In sum, Professor Waterman is a national treasure! If I can add a personal thought, he brought us closer to von Neumann. Here are the lessons I learned from him, stated as ''axioms,'' which are shared lessons from von Neumann.

Michael Waterman's axioms:

- Axiom 1. Intertwining continuous and discrete mathematics is quintessential to computational biology.
- Axiom 2. Be a world class scientist in at least one of the areas of your interdisciplinary research.
- Axiom 3. Be an intra-math, intersciences, and cross-cultures scientist.
- Axiom 4. Be guardian of mathematical rigor.
- Axiom 5. If you can't say something good about someone, don't say anything at all.
- Axiom 6. You can know a person by his/her heroes.
- Axiom 7. And in the end, the love you take is equal to the love you make.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

*https://bio.tools/atac

REFERENCES

Andersen, J.E., Chekhov, L.O., Penner, R., et al. 2013. Topological recursion for chord diagrams, RNA complexes, and cells in moduli spaces. *Nucl. Phys. B.* 866, 414–443.

Arratia, R., Goldstein, L., and Gordon, L. 1990a. Poisson approximation and the chen-stein method. *Stat. Sci*. 5, 403–424.

Arratia, R., Gordon, L., and Waterman, M.S. 1990b. The erdos-rényi law in distribution for coin tossing and sequence matching. *Ann. Stat*. 18, 539–570.

Arratia, R., Lander, E.S., Tavaré, S., et al. 1991. Genomic mapping by anchoring random clones: A mathematical analysis. *Genomics* 11, 806–827.

Arratia, R., and Waterman, M.S. 1985. An erdös-rényi law with shifts. *Adv. Math*. 55, 13–23.

Arratia, R., and Waterman, M.S. 1994. A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab*. 4, 200–225.

Barrett, C., He, Q., Huang, F.W., et al. 2019. A Boltzmann sampler for 1-pairs with double filtration. *J. Comput. Biol*. 26, 173–192.

Breen, S., Waterman, M.S., and Zhang, N. 1985. Renewal theory for several patterns. *J. Appl. Probab*. 22, 228–234.

Churchill, G.A., and Waterman, M.S. 1992. The accuracy of DNA sequences: Estimating sequence quality. *Genomics* 14, 89–98.

Deonier, R.C., Tavaré, S., and Waterman, M.S. 2005. *Computational Genome Analysis: An Introduction*. Springer Science & Business Media, New York, NY, USA.

Drmanac, R., Labat, I., Brukner, I., et al. 1989. Sequencing of megabase plus DNA by hybridization: Theory of the method. *Genomics*. 4, 114–128.

Fresco, J.R., Alberts, B.M., Doty, P., et al. 1960. Some molecular details of the secondary structure of ribonucleic acid. *Nature* 188, 98–101.

Galas, D.J., Eggert, M., and Waterman, M.S. 1985. Rigorous pattern-recognition methods for DNA sequences: Analysis of promoter sequences from Escherichia coli. *J. Mol. Biol*. 186, 117–128.

Goldstein, L., and Waterman, M.S. 1992. Poisson, compound poisson and process approximations for testing statistical significance in sequence comparisons. *Bull. Math. Biol*. 54, 785–812.

Gordon, L., Schilling, M.F., and Waterman, M.S. 1986. An extreme value theory for long head runs. *Probab. Theory. Relat. Fields* 72, 279–287.

Howell, J., Smith, T., and Waterman, M. 1980. Computation of generating functions for biological molecules. *SIAM J. Appl. Math*. 39, 119–133.

Huang, F.W., and Reidys, C.M. 2016. Topological language for RNA. *Math. Biosci.* 282, 109–120.

Idury, R., and Waterman, M. 1995. A new algorithm for DNA sequence assembly. *J. Comput. Biol*. 2, 291–306.

Istrail, S. 2016. Eric Davidson: Master of the Universe. *Dev. Bio*. 412:S47–S54.

Istrail, S., De-Leon, S.B.-T., and Davidson, E.H. 2007. The regulatory genome and the computer. *Dev. Bio.* 310, 187–195.

Istrail, S., Sutton, G., Florea, L., et al. 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Nat. Acad. Sci.* 101, 1916–1921.

Kim, J.H., Waterman, M.S., and Li, L.M. 2007. Accuracy assessment of diploid consensus sequences. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 4, 88–97.

Knuth, D. 1970. Permutations, matrices, and generalized young tableaux. *Pac. J. Math*. 34, 709–727.

Lander, E.S., and Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2, 231–239.

Li, L.M., Kim, J.H., and Waterman, M.S. 2004. Haplotype reconstruction from snp alignment. *J. Comput. Biol*. 11, 505–516.

Lippert, R.A., Huang, H., and Waterman, M.S. 2002. Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl. Acad. Sci. U. S. A*. 99, 13980–13989.

Liu, X., Wan, L., Li, J., et al. 2011. New powerful statistics for alignment-free sequence comparison under a pattern transfer model. *J. Theor. Biol.* 284, 106–116.

McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29, 1105–1119.

Penner, R., and Waterman, M.S. 1993. Spaces of RNA secondary structures. *Adv. Math*. 101, 31–49.

Pevzner, P. 1989. *L*-tuple DNA sequencing: Computer analysis. *J. Biomol. Struct. Dyn*. 7, 63–73.

Pevzner, P.A., Tang, H., and Waterman, M.S. 2001. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A*. 98, 9748–9753.

Port, E., Sun, F., Martin, D., and Waterman, M.S. 1995. Genomic mapping by end-characterized random clones: A mathematical analysis. *Genomics* 26, 84–100.

Reinert, G., Chew, D., Sun, F., et al. 2009. Alignment-free sequence comparison (I): Statistics and power. *J. Comput. Biol*. 16, 1615–1634.

Reinert, G., Schbath, S., and Waterman, M.S. 2000. Probabilistic and statistical properties of words: An overview. *J. Comput. Biol.* 7, 1–46.

Ren, J., Bai, X., Lu, Y.Y., et al. 2018. Alignment-free sequence analysis and applications. *Annu. Rev. Biomed. Data Sci*. 1, 93–114.

Robinson, G. 1938. On the representations of the symmetric group. *Am. J. Math.* 60, 745–760.

Sankoff, D. 2000. The early introduction of dynamic programming into computational biology. *Bioinformatics* 16, 41–47.

Schensted, C. 1961. Longest increasing and decreasing subsequences. *Canad. J. Math*. 13, 179–191.

Schmitt, W.R., and Waterman, M.S. 1994. Linear trees and RNA secondary structure. *Discrete. Appl. Math*. 51, 317–323.

Sellers, P. 1980. The theory and computation of evolutionary distances: Pattern recognition. *J. Algorithm* 1, 359–373.

Smith, T.F., and Waterman, M.S. 1981a. Comparison of biosequences. *Adv. Appl. Math*. 2, 482–489.

Smith, T.F., and Waterman, M.S. 1981b. The identification of common molecular subsequences. *J. Mol. Biol*. 147, 195–197.

Smith, T.F., Waterman, M.S., and Burks, C. 1985. The statistical distribution of nucleic acid similarities. *Nucleic Acids Res*. 13, 645–656.

Smith, T.F., Waterman, M.S., and Fitch, W.M. 1981. Comparative biosequence metrics. *J. Mol. Evol*. 18, 38–46.

Song, K., Ren, J., Reinert, G., et al. 2014. New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing. *Brief. Bioinform*. 15, 343–353.

Stein, P.R., and Waterman, M.S. 1979. On some new sequences generalizing the catalan and motzkin numbers. *Discrete Math*. 26, 261–272.

Tinoco, I., Uhlenbeck, O.C., and Levine, M.D. 1971. Estimation of secondary structure in ribonucleic acids. *Nature* 230, 362.

Torney, D.C., Burks, C., Davison, D., et al. 1990. Computation of d2: A measure of sequence dissimilarity, 109–125. *In Computers and DNA*, The Proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop, December 12–16, 1988, Routledge, Santa Fe, New Mexico, USA.

Vingron, M., and Waterman, M.S. 1994. Sequence alignment and penalty choice: Review of concepts, case studies and implications. *J. Mol. Biol*. 235, 1–12.

Wan, L., Reinert, G., Sun, F., and Waterman, M.S. 2010. Alignment-free sequence comparison (II): Theoretical power of comparison statistics. *J. Comput. Biol*. 17, 1467–1490.

Waterman, M., Arratia, R., and Galas, D. 1984. Pattern recognition in several sequences: Consensus and alignment. *Bull. Math. Biol*. 46, 515–527.

Waterman, M.S. 1978. Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Stud*. 1, 167–212.

Waterman, M.S. 1979. Combinatorics of RNA hairpins and cloverleaves. *Stud. Appl. Math*. 60, 91–98.

Waterman, M.S. 1983. Frequencies of restriction sites. *Nucleic Acids Res*. 11, 8951–8956.

Waterman, M.S. 1989a. Consensus methods for folding single-stranded nucleic acids. *In* Waterman, M.S., ed. *Mathematical Methods for DNA Sequences/Editor*. CRC Press, London, UK.

Waterman, M.S. 1989b. *Mathematical Methods for DNA Sequences*. CRC Press, Inc.: Boca Raton, FL, USA.

Waterman, M.S. 1995. *Introduction to Computational Biology*. Chapman and Hall, London.

Waterman, M.S. 2016. *Getting Outside: A Far-Western Childhood*. CreateSpace Independent Publishing Platform.

Waterman, M.S., and Eggert, M. 1987. A new algorithm for best subsequence alignments with application to trna-rrna comparisons. *J. Mol. Biol*. 197, 723–728.

Waterman, M.S., Eggert, M., and Lander, E. 1992. Parametric sequence comparisons. *Proc. Natl. Acad. Sci. U. S. A*. 89, 6090–6093.

Waterman, M.S., Gordon, L., and Arratia, R. 1987. Phase transitions in sequence matches and nucleic acid structure. *Proc. Natl. Acad. Sci. U. S. A*. 84, 1239–1243.

Waterman, M.S., and Perlwitz, M.D. 1984. Line geometries for sequence comparisons. *Bull. Math. Biol*. 46, 567–577.

Waterman, M.S., and Smith, T.F. 1978. RNA secondary structure: A complete mathematical analysis. *Math. Biosci*. 42, 257–266.

Waterman, M.S., and Smith, T.F. 1986. Rapid dynamic programming algorithms for RNA secondary structure. *Adv. Appl. Math*. 7, 455–464.

Waterman, M.S., and Vingron, M. 1994. Sequence comparison significance and poisson approximation. *Stat. Sci*. 9, 367–381.

Zhang, K., Calabrese, P., Nordborg, M., et al. 2002a. Haplotype block structure and its applications to association studies: Power and study designs. *Am. J. Hum. Genet*. 71, 1386–1394.

Zhang, K., Deng, M., Chen, T., et al. 2002b. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. U. S. A*. 99, 7335–7339.

Zhang, K., Qin, Z., Chen, T., et al. 2005. Hapblock: Haplotype block partitioning and tag snp selection software using a set of dynamic programming algorithms. *Bioinformatics* 21, 131–134.

Zhang, K., Qin, Z.S., Liu, J.S., et al. 2004. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res*. 14, 908–916.

Zhang, K., Sun, F., Waterman, M.S., et al. 2003. Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am. J. Hum. Genet*. 73, 63–73.

Address correspondence to:
Fengzhu Sun, PhD
Department of Quantitative and Computational Biology
University of Southern California
Los Angeles, CA 90089
USA

E-mail: fsun@usc.edu