

HPS DNA DATA MINING SYSTEM

(c) 2007 [HPS Transform Research](#)
by [Nelson R. Manohar, Ph.D.](#)
nelsonmanohar@yahoo.com

FOREWORD

THIS DOCUMENT PRESENTS PRELIMINARY PATENT-PENDING RESEARCH RESULTS; THEREFORE DISCLOSURE LEVEL IS PURPOSELY CURTAILED AND LIMITED. The findings presented here relate to a theoretical signal processing and data mining breakthrough, demonstrated herewith through its application to fundamental similarity/homology-based problems on the bioinformatics domain area. These systems (herein referred to as "HPS DNA mining systems") on their own right constitute a significant and evolving body of bioinformatics research contributions and deliverables. The long-term orientation for this exploratory interdisciplinary research is directed towards enabling large-scale comparative genomics research.

AUTO-GENERATED REPORT FOR SIMILARITY SEARCH

This **auto-generated** report presents results of a **sequence aligning and phylogenetic mining search**. Given a DNA signature and a DNA test sequence, "HPS DNA mining systems" are capable of correctly identifying, extracting, and aligning ALL (that is, zero or more) TRUE matching instances of said DNA signature found to be within said DNA test sequence. We define a **true matching instance** to be a subsequence of the DNA test sequence for which either of the following (necessary and sufficient) criteria is met: (1) **the subsequence is an EXACT replica (100% compatible) of the DNA signature** or (2) **the subsequence exhibits TOLERABLE accumulation of DNA match errors WHILE retaining certain identifying qualities**. Specifically, accumulation of error may take place in any of the following ways: (2a) a matching instance exhibits a tolerable accumulation of **SPORADIC SINGLE-POINT BASE-PAIR MUTATIONS** (such as inversions, transcription errors, duplications, etc.), (2b) a matching instance exhibits tolerable error bursts of **DELETED base-pairs** in either DNA signature or DNA test sequence, (2c) a matching instance exhibits tolerable error bursts of **INSERTED base-pairs** in either DNA signature or the DNA test sequence, or (2d) a matching instance exhibits tolerable accumulation of error due to **ANY COMBINATION OF THE ABOVE**. The details of how data mining parameters are selected by "HPS DNA mining systems" are purposely curtailed. Suffice to say that the user needs only provide the DNA signature and the DNA test sequence and nothing more. "HPS DNA mining systems" will then examine potential matching instances of the DNA signature found within the DNA test sequence and reports true, optimal, and feasible matching instances found within.

NEW RUN-TIME COMPLEXITY LOWER BOUNDS

More importantly, "HPS DNA mining systems" complete said tasks within NEW AND LOWER RUN-TIME COMPLEXITY BOUNDS previously considered not feasible in practical cases and general use. For example, "HPS DNA mining systems" can find matching instances of a DNA signature of size **M** within a sequence of size **N** in sub-linear time with respect to the input size **N**. That is, **"HPS DNA mining systems" can perform (certain types of) COMBINATORIAL DATA MINING over sequences of size N in SUB-LINEAR TIME!** As a matter of fact, in practice (and regardless of input size), linear run-time cost operations (such as time-series reading/writing) do take more time than the run-time cost for the "HPS DNA mining systems" data mining core which (as stated) performs a COMBINATORIAL ANALYSIS of the entire DNA sequence. Those versed in the arts, will recognize this to be an extraordinary result of vast implications.

PHYLOGENETIC AND MUTATION SEARCH

Moreover, **given R potentially phylogenetic DNA sequences**, by simply selecting (subsequence content) from any one such as the DNA signature, "HPS DNA mining systems" make possible to **find true matching instances from the remaining R-1** (potentially phylogenetic or damaged) DNA sequences in optimal time. Matching instances unearthed by "HPS DNA mining systems" are guaranteed to be true instances of the given DNA signature, this being true EVEN when said matching instances may exhibit substantial DNA damage (for example, due to mutations, deletions, inversions, insertions, duplications,

transcription errors). Those versed in the arts will recognize that such error cases typically complicate correct identification by inducing probabilities of misidentifications as well as resulting in significantly larger (such as sub-quadratic as opposed to sub-linear) run-time computational costs in existing alternative approaches in use today. Moreover, "HPS DNA mining systems" not only unearth and identify true matching instances but also produce rich and detailed output which identifies (1) all **DIFFERENCES** between a DNA signature and any such matching instance (2) as well as all **REPAIRS** (e.g., the logical placement of necessary DELETES and/or INSERTS operations) needed to transform (or repair) any (**potentially phylogenetic or damaged**) matching instance into the DNA signature (and vice versa). For example, "HPS DNA mining systems" produce alignment-edit graphs, similarity graphs, base-to-base comparison reports, graphical identification of matching instances, and other not yet disclosed analysis reports and graphs, etc.

DATABASE AND WEB MINING

Additionally, "HPS DNA mining systems" allow the search for matching instances of the given DNA signature to be performed against a DNA sequence database (whether such is remotely located (as in the case of ENTREZ) or stored on your hard-disk. Moreover, database sequences can be accessed in either the ENTREZ format or in a time series format ((genomic-address, base) tuples). Moreover, "HPS DNA mining systems" are capable of automatically extracting time-series data from DNA sequence data found in ENTREZ webpages.

DEMONSTRATION DATABASE

For demonstration purposes, "HPS DNA mining systems" have been tested on the complete Escherichia coli K12 bacterium genome (that is, a genome of approximately 4.7 million base pairs (bp) and close to 400 DNA (10,000+ bp) DNA fragments). The Escherichia coli K12 bacteria genome (U00096) - by *Blattner, F.R., Plunkett, G. III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y., from "The complete genome sequence of Escherichia coli K-12", Science 277 (5331), 1453-1474 (1997)* - was obtained from ENTREZ.

DNA INPUTS: SIGNATURE AND TEST SEQUENCE

The DNA sequences you provided for us to analyze were contained within TWO input data files (these corresponding to DNA SIGNATURE and DNA TEST SEQUENCE). These two DNA sequences were automatically merged into a global sequence datafile, which was used for internal representation of the inputs. No changes to your original data were done UNLESS you explicitly specified the application of a test mutation to one of the two sequences AND requested that the resultant mutation be stored into the corresponding input file. However, a result of this scheme is that genomic addresses shown on this SUMMARY REPORT page are SHIFTED by the (ACGT-based) size of the DNA signature. To find the true and exacting genomic addresses (with respect to your given DNA test sequence), please access the link to the DETAILED REPORT page associated with any of the resultant matching instance. The DNA sequences analyzed were the following:

DNA_INPUT_SEQUENCE	NUMBER_OF_BASES	DESCRIPTION_OF_THE_DNA_INPUT_SEQUENCE	FILENAME_CONTAINING_THE_INPUT_SEQUENCE
DNA_SIGNATURE	1000	1DNA-BASES-E-COLI-K12-1786520-FROM5000-TO5999	F:/HPS_DATA/HPS_INPUTS/HPS_1DNA_SIGNATURE_BASES.DNA
DNA TEST SEQUENCE	32977	1DNA-BASES-SEQ1786520TESTCASE	F:/HPS_DATA/HPS_INPUTS/HPS_1DNA_TESTSEQUENCE_BASES.DNA

DESCRIPTION OF THE TEST DATA

THIS SECTION WAS NOT AUTO-GENERATED. It is provided for convenience of those seeking to review, rate, or compare the results presented herein. To test the HPS DNA pattern miner, a DNA input sequence of roughly 20000 base pairs (bp) was constructed using the 13000 bp E-COLI DNA sequence 1786520 (obtained from ENTREZ) as a starting point. Said sequence was then MODIFIED as follows. First, the DNA signature was selected to be simply the bases contained within the genomic addresses 5000 TO 5999. Then, to test several key conditions, various **damaged instances of the DNA signature** were manufactured (as

described below) and INSERTED INTO SPECIFIC LOCATIONS. This augmentation process is described below. AS SHOWN, ALL TEST CASES WERE CORRECTLY UNEARTHED, RATED, AND REPORTED.

LAYOUT OF TEST-CASE MATCHING INSTANCES

Between [5000:6000]	the identical matching instance of the DNA signature is found.	The original DNA test sequence 1786520 had 13480 bp.
Between [7000:8000]	a heavily single-point mutated version of the original DNA signature was inserted, consisting of approximately pseudo-random 25 single-point mutations from A->T coupled with approximately 25 single-point mutations from T->A. The mutations were chosen to take place at particular indexes (e.g., every 10th base which happened to be A) to easily verify via eye inspection of proper detection within the matching instance. Effectively, the resulting matching instance had about 5% (50/1000) randomly single-point mutations which HPS algorithms would have to deal with in the identification of the intrinsic structure of the underlying DNA signature found within.	This transformation increased the resultant DNA test sequence in size from 13480 to 14480.
Between [13481:14450]	a heavily damaged version of the DNA signature was inserted, which exhibited a SINGLE DELETION BURST of exactly 30 base pairs at (intra) genomic address (800). To accomplish this, a true DNA signature match was inserted therein, but then basepairs at (DNA signature) genomic addresses [800:830] were deleted. No single point mutations were inflicted onto this true matching instance.	This increased the resultant DNA test sequence in size from 14480 to 15550.
Between [14550:17000]	a DNA filler sequence extracted from a subsequence of the original DNA test sequence (but OUTSIDE of the DNA signature region) was inserted into the DNA test sequence.	This increased the size from 15550 to 17000.
Between [17000:18000]	a heavily damaged version of the DNA signature was inserted, which exhibited (1) a DELETION BURST of exactly 30 base pairs at (intra) genomic address (300) coupled with (2) a RANDOM INSERTION BURST of 30 base pairs at (intra) genomic address (800). Effectively, a true DNA signature match was inserted therein, but with deleted basepairs at genomic addresses [300:330] and inserted RANDOM bases at genomic addresses [800:830]. No single point mutations were inflicted onto this true matching instance.	This increased the size from 17000 to 18000.
Between [18000:19000]	another DNA filler sequence extracted from a subsequence of the original DNA test sequence (but OUTSIDE of the DNA signature region) was inserted into the DNA test sequence.	This increased the size from 18000 to 19000.
Between [19000:19970]	a heavily damaged version of the DNA signature was inserted, which exhibited a SINGLE DELETION BURST of exactly 30 base pairs at (intra) genomic address (300). Effectively, a true DNA signature match was inserted therein, but with deleted basepairs at genomic addresses [300:330]. No single point mutations were inflicted onto this true matching instance.	This increased the size from 19000 to 20000.
Between [19970:30050]	a DNA filler sequence extracted from SEQUENCE_1786454.DNA (10080 bp) was inserted.	This increased the size from 20000 to 30000.
Between [24000:24975]	a heavily damaged version of the DNA signature was inserted, which exhibited (1) multiple DELETION BURSTS of exactly 8 base pairs at every 200 (intra) genomic address (i.e., 200, 400, 600, 800) and then coupled with (2) a single INSERTION BURST of 8 base pairs at (intra) genomic address (500). No single point mutations were inflicted onto this true matching instance.	This increased the size from 30000 to 31000.
Between [30000:30975]	a heavily damaged version of the DNA signature was inserted, which exhibited the above DNA damage coupled with several random single point mutations at periodic intervals (1 base every 50 bases, for the first 500 genomic addresses).	This increased the size from 31000 to 32000.
Between [32002:32978]	a heavily damaged version of the DNA signature was inserted, which exhibited multiple DELETION BURSTS of exactly 8 base pairs at every 200 (intra) genomic address (i.e., 200, 400, 600, 800).	This increased the size from 32000 to 33000. Total input size (approx. 1000+33000bp).

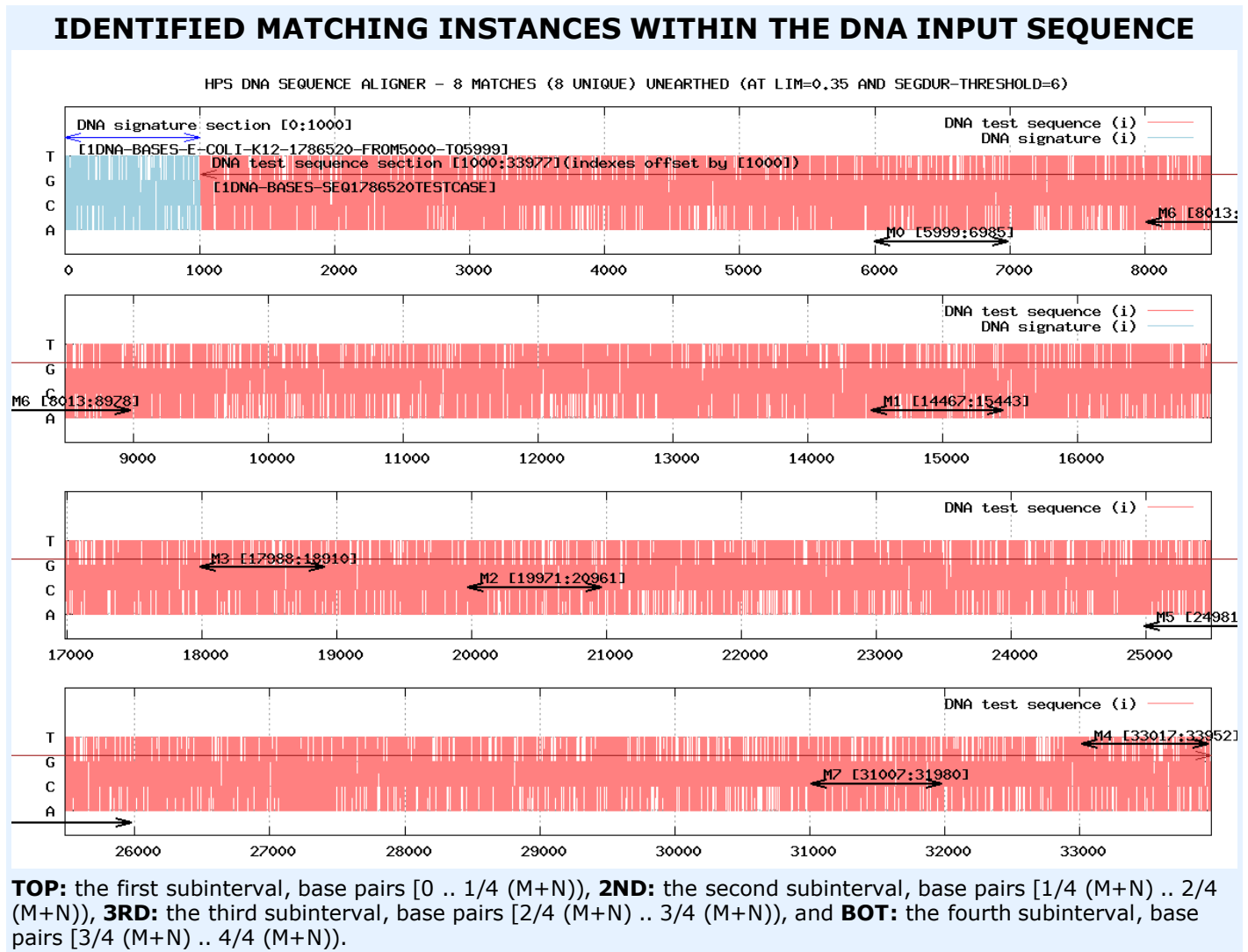
IDENTIFIED MATCHING INSTANCES OF DNA SIGNATURE

The following figure identifies the approximate relative placement of matching instances of the DNA signature found to exist within the given DNA test sequence. To enhance viewing and close up examination of findings, the resulting time plot has been divided into four close-up subintervals of approximately the same size. Correspondingly, from top to bottom, a time-plot panel shows each such subinterval, with the top panel containing the first quarter of the span of the DNA input sequence and the bottom panel containing the last

quarter. As shown, the (combined) DNA input sequence consists of the DNA signature being prefixed to the DNA test sequence. Therefore, genomic addresses shown are shifted by the size of the DNA signature.

IDENTIFICATION OF MATCHING INSTANCES

Matching instances of the DNA signature are identified via a **SOLID BLACK ARROW**. Each such arrow is also annotated with a TEXT LABEL having the format: "**M# [START:END]**"; where **M#** uniquely identifies a matching instance across all auto-generated "HPS DNA mining systems" reports. For example, for any matching instance on the SUMMARY REPORT page, there exists a DETAILED REPORT page, identified and linked by said number. **START** and **END** represent the starting and ending genomic addresses of the matching instance as found within the (combined) DNA input sequence. These genomic addresses are APPROXIMATE. To obtain the TRUE AND EXACTING genomic address for any matching instance with respect to the given DNA test sequence, you MUST refer to said matching instance's DETAILED REPORT page. Nevertheless, the TRUE genomic address can also be estimated by simply subtracting the size of the DNA signature from such approximate alignment indexes. However, this number represents only an approximation that lies within +/- a small constant.



RECONSTRUCTION OF BURST DAMAGED DNA (OR PHYLOGENETIC) MATCHING INSTANCES

Matching instances unearthed by "HPS DNA mining systems" may actually be (mutated as well as potentially phylogenetic) versions of the given DNA signature, which nevertheless retain some special intrinsic qualities associated with the DNA signature. Typically, in such cases, special operations are used to show how, for

any matching instance said to represent a damaged or mutated version of the DNA signature, it is possible to (optimally) reconstruct (i.e. with few operations and few changes) the targeted DNA signature sequence.

RECONSTRUCTION OPERATIONS

The minimal abstract base machine needed to perform such reconstruction consists of just two operations: **DELETE(X)** and **INSERT(X)**. The **DELETE(X)** operation edits a DNA sequence (at the current genomic address) by deleting the next **X bases** from it. The **INSERT(X)** operation edits a DNA sequence (at the current genomic address) by inserting **X base placeholders** into it. A **DELETE/INSERT** operation is triggered when a damage burst of more than **Z bases** is observed within a DNA sequence. Suppose that such event is detected at genomic address **Y**. Then, such triggering event will be identified by the **RESYNC (Y, Z)** operation, which is therefore always followed by either a **DELETE(X)** or **INSERT(X)** operation. Finally, we choose to use another couple of operations to identify the exacting genomic address of any matching instance. First, a **PRELIM-ALIGN(Y)** operation marks the preliminary genomic address believed to be the start of a matching instance. Then, an **OFFSET-ALIGN(W)** operation identifies how far away (in number of bases, in either direction) is the exacting genomic address that corresponds to the start of the matching instance. Note that (as stated), both these genomic addresses refer to indexing within the DNA input sequence, which consists of the DNA signature PREPENDED to the DNA test sequence). As stated, to obtain the true genomic address (MSTART) for the start of a matching instance, one need only to refer to the DETAILED REPORT page for said matching instance or alternatively, subtract the size **M** of the DNA signature from the offset-corrected preliminary addresses. That is, **MSTART=Y+W-M**.

ADJUSTED AND ORIGINAL PAIRINGS

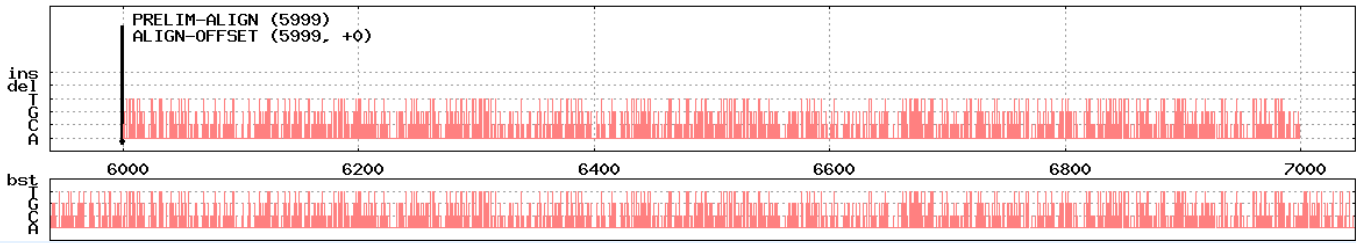
Finally, to differentiate between original and reconstructed cases, we refer to the pairing of the DNA signature to the ORIGINAL (i.e., as found) matching instance M# from the DNA test sequence to as the **ORIGINAL SIG-TO-M# pairing**. Similarly, we refer to the pairing of the DNA signature to the ADJUSTED (i.e., after reconstruction operations) matching instance M# to as the **ADJUSTED SIG-TO-M# pairing**. The following figures show, for each matching instance extracted from the DNA test sequence, the resultant reconstruction operations needed to recover the DNA signature from said matching instance. For this reason, each figure has two parts. Whereas the top part of each figure shows the **ADJUSTED SIG-TO-M# pairing** together with the various reconstruction operations needed, the bottom part of each figure shows the **ORIGINAL SIG-TO-M# pairing** of DNA signature to matching instance (as given). Within any pairing plot, the DNA signature is always colored in "light-blue" color whereas the matching instance is always colored in "light-pink".

BEHAVIOR OF THE RECONSTRUCTION

Note that when a matching instance exhibits no burst DNA damage, no reconstruction needs to be apply (i.e., all DNA damage is due to single-point errors) and thus, ORIGINAL and ADJUSTED pairings are the same. Moreover, note that when a matching instance exhibits burst DNA damage, while the ORIGINAL pairing shows clear cross-correspondence discrepancies between DNA signature and matching instance due to the misalignment induced by such burst DNA damage, after such burst DNA damage is corrected via reconstruction operations, the resultant ADJUSTED pairing exhibits consistently accurate cross-correspondence. Note however, that while burst DNA damage is corrected during reconstruction, single point random errors are always left (on either pairing) uncorrected (such showing in various test-cases). To this end, note that for each matching instance a summary statistics tuple is also given, which shows the number of mismatched base pairs in the cross-correspondence check between the DNA signature and the ADJUSTED (reconstructed) matching instance. Such reported mismatches are due to single-point random errors plus the repair cost (in bases) of reconstruction operations.

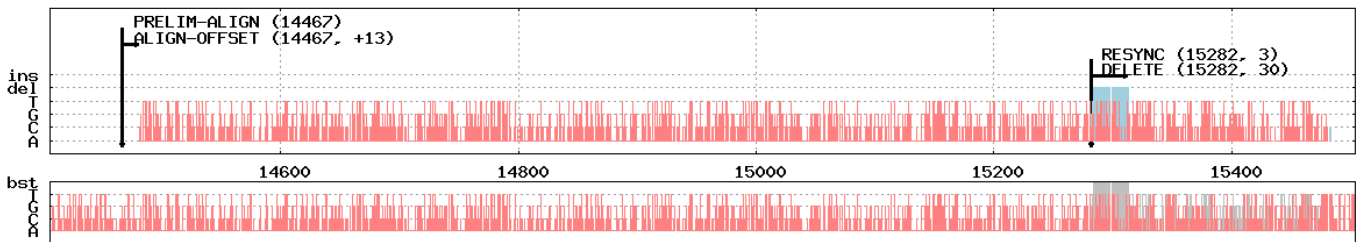
M0: DETAILED REPORT FOR MATCHING INSTANCE

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	NONMATCHBP
5999	6985	0.73170733	987	906.0	0



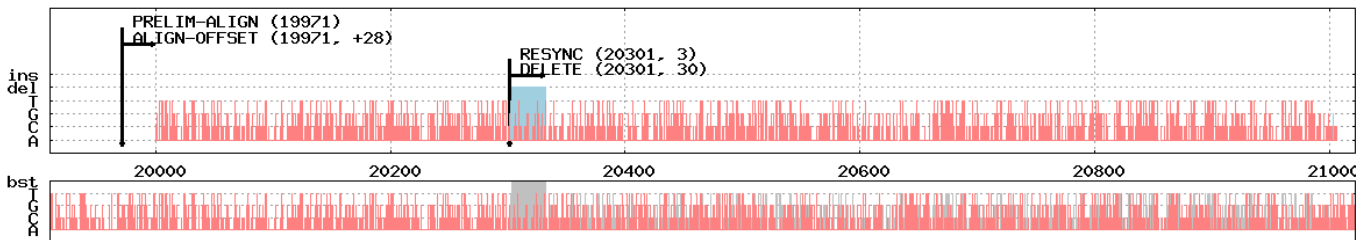
M1: DETAILED REPORT FOR MATCHING INSTANCE

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	NONMATCHBP
14467	15443	0.61538464	977	1221.0	34



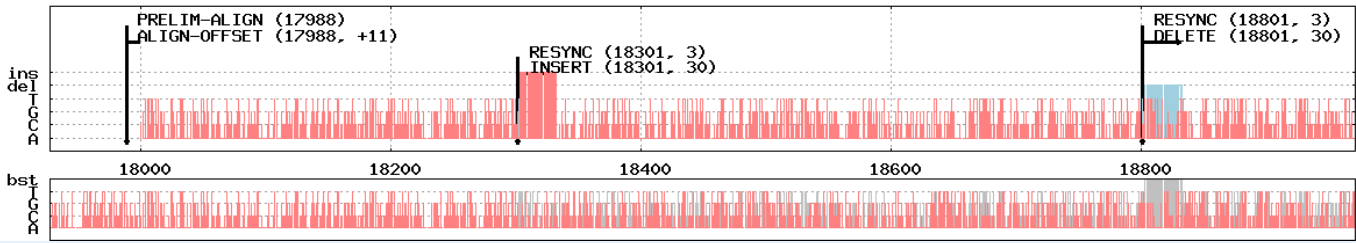
M2: DETAILED REPORT FOR MATCHING INSTANCE

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	NONMATCHBP
19971	20961	0.5609756	991	1272.0	34



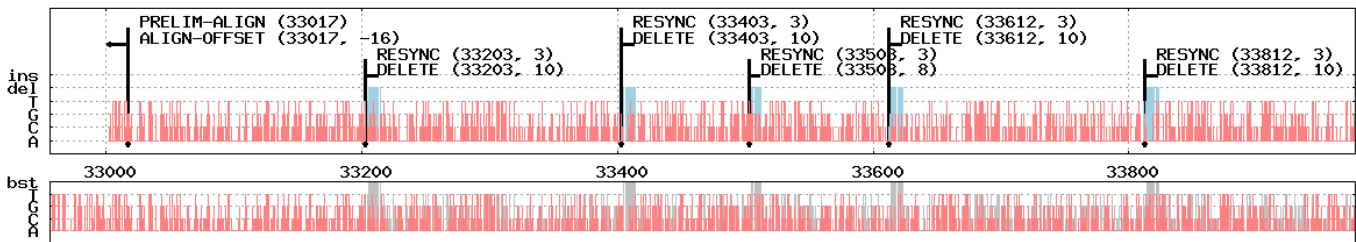
M3: DETAILED REPORT FOR MATCHING INSTANCE

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	NONMATCHBP
17988	18910	0.45	923	1268.0	64



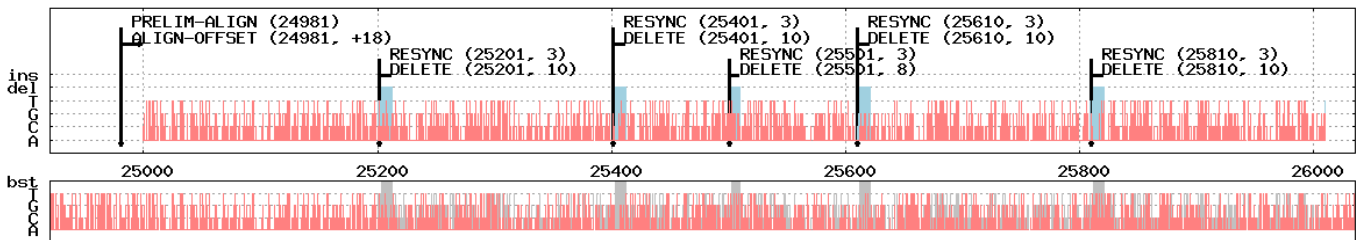
M4: DETAILED REPORT FOR MATCHING INSTANCE

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	NONMATCHBP
33017	33952	0.44444445	936	1416.0	68



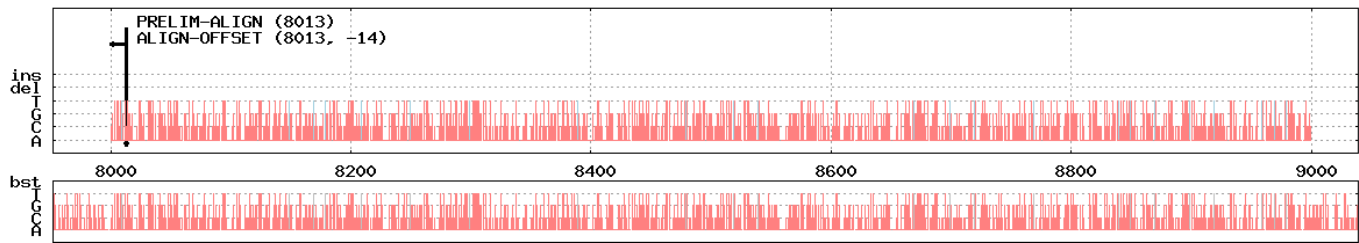
M5: DETAILED REPORT FOR MATCHING INSTANCE

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	NONMATCHBP
24981	25975	0.41463414	995	1603.0	68



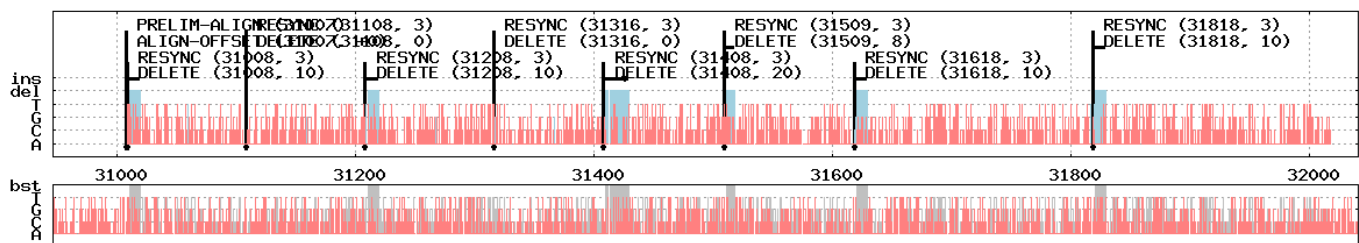
M6: DETAILED REPORT FOR MATCHING INSTANCE

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	NONMATCHBP
8013	8978	0.3902439	966	1706.0	42



M7: DETAILED REPORT FOR MATCHING INSTANCE

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	NONMATCHBP
31007	31980	0.375	974	1272.0	106

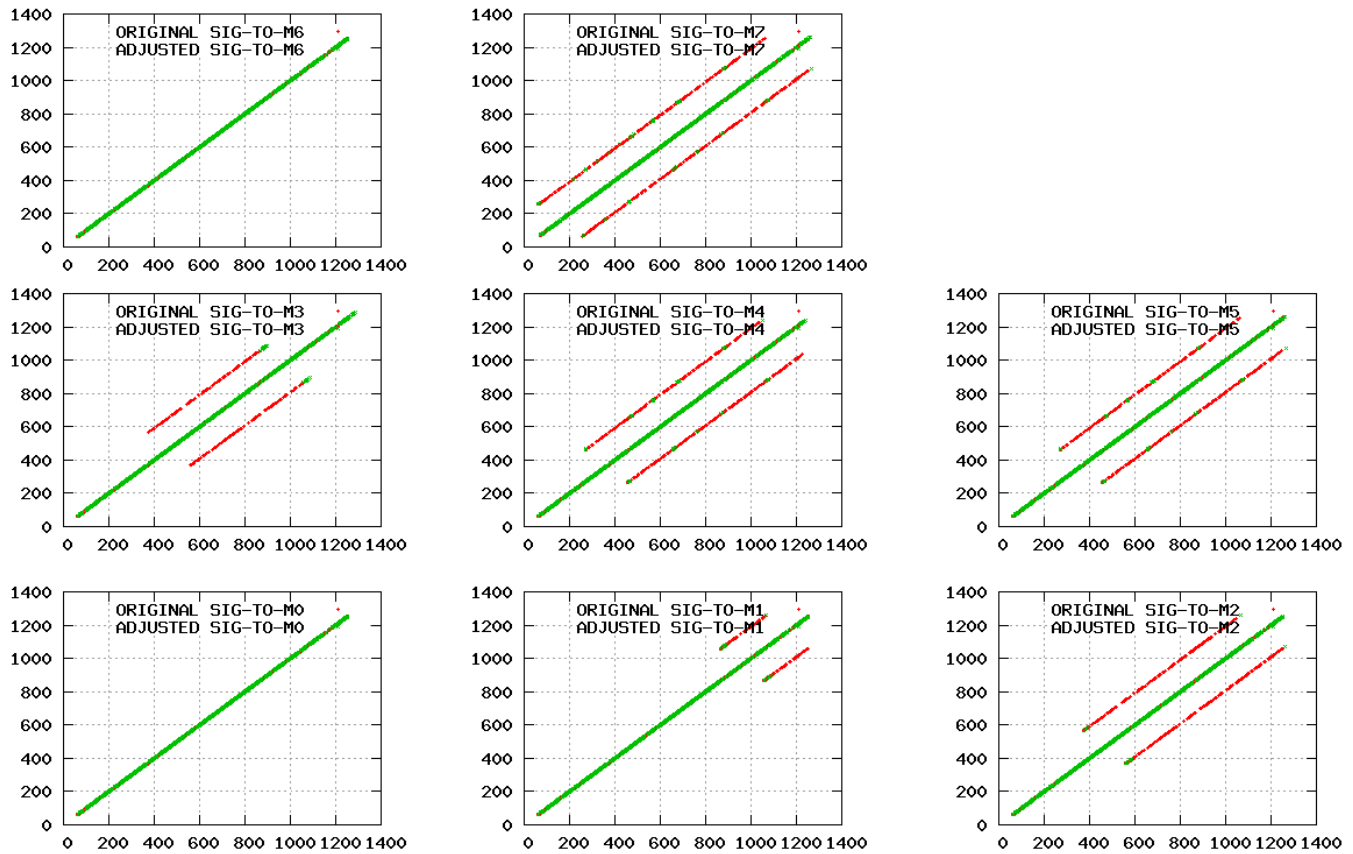


HPS DNA mining systems (c) 2005, 2006, 2007 - Dr. Nelson R. Manohar. All rights reserved.

SIMILARITY PLOTS BETWEEN DNA SIGNATURE AND MATCHING INSTANCES

Similarity plots (i.e., dot-matrix alignment plots) provide quick visual identification of the fitness of the resultant alignment between the DNA signature and a given matching instance. Note that TWO similarity plots are shown on each panel, a RED line and a GREEN line. The RED one corresponds to a similarity plot between the DNA signature and the ORIGINAL DNA subsequence data that corresponds to the matching instance found genomic addresses. The GREEN line represents a similarity plot between the DNA signature and the ADJUSTED DNA subsequence obtained through the application of RECONSTRUCTION operations over the ORIGINAL DNA subsequence. Recall that reconstruction operations are used to correct BURSTS of DNA damage found to be present within the ORIGINAL DNA subsequence. Such reconstructions generate a new ADJUSTED outlook over the ORIGINAL the DNA subsequence data of the matching instance, on which DNA bases or placeholders are deleted or inserted to maintain synchrony. Therefore, note how GREEN and RED plots DO differ in cases where a matching instance exhibits BURST DNA damage. In such cases, note that the GREEN line exhibits FAITHFUL ALIGNMENT (just after reconstruction operations are applied). Such case manifests, during each burst of DNA damage, as a reset of the green line back to the X=Y diagonal just after a COUPLE OF INITIAL BURST-DETECTION MISSES. These initial mismatches are due to an initial detection cost and relate to algorithmic cost decisions. That is, the cost of reconstruction operations manifest as initial short-term GREEN discrepancies that always coincide with the start of large RED discrepancies and that after such initial repair cost, the GREEN line depicts through the resynching of the ADJUSTED matching instance to the DNA signature, whereas the RED similarity plot depicts a furtherance of the discrepancy between the ORIGINAL matching instance with respect to the DNA signature. The detailed analysis of such resynching is found in the DETAILED REPORT page that corresponds to said matching instance. As stated, only **TRUE, OPTIMAL, and FEASIBLE matching instances** are reported. Finally, plots are generated using an IN-HOUSE distance metric designed for visual acuity, which somewhat distorts the genomic address correspondence of those points lying OUTSIDE the X=Y line (i.e., the discrepancies).

"SIMILARITY PLOTS" DNA SIGNATURE AGAINST MATCHING INSTANCE (UNDER UNADJUSTED ALIGNMENT)



FROM BOTTOM LEFT TO TOP RIGHT: Similarity plots (TWO per panel). The RED one is a similarity plot between the DNA signature and the ORIGINAL DNA subsequence corresponding to the matching instance found. The GREEN one is a similarity plot between the DNA signature and the ADJUSTED DNA subsequence obtained through the application of reconstruction operations.

DNA SIGNATURE-ROOTED PHYLOGRAM

As stated, all matching instances are TRUE matching instances of the DNA signature, i.e., the limiting probability of (such matching instance being a TRUE match to the DNA signature) is ONE in the formal sense (i.e., virtual certainty for all practical purposes). Therefore, the distance metrics used in this plot rather than representing measurements of the probabilistic fitness of any such match represent instead abstraction measures of the DNA damage that is accounted for within a matching instance. As shown above, in the reconstruction edit plots, it has been determined that after such specified DNA reconstruction, all such burst DNA damage could be corrected and an exacting reproduction of the original DNA signature is therefore achievable from the matching instance.

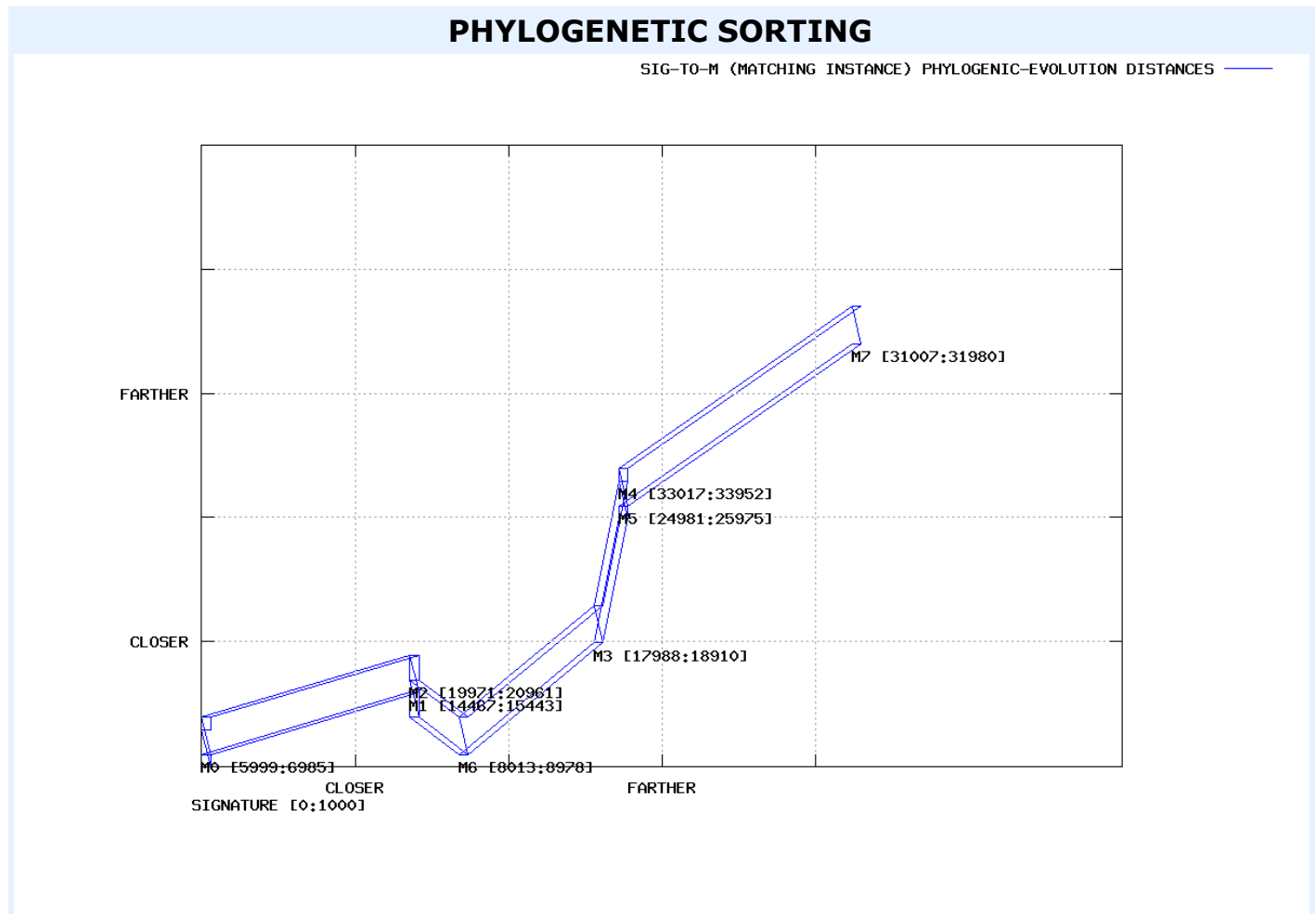
RELATIVE EVOLUTIONARY COST METRICS

In the following plot, points represent the exacting (x,y,z) cost-metric positioning of the given matching instance, The line shown traverses the set of matching instances along a sorted partial order based on the total number of uncorrectable errors (as well as some other cost-metrics). This way, the DNA signature is found at the $(0,0,0)$ coordinates and each point along the line represents a matching instance of greater number of accumulated uncorrectable errors. In actuality, the evolutionary cost-metric used to generate the plot is a three-dimensional cost metric and thus this ordering is a partial order in the formal sense. As a result, the positioning (x_2, y_2, z_2) of the subsequent (partially ordered) matching instance is found to be in (and displayed as) a proportional delta with respect to the previous (x_1, y_1, z_1) coordinates. This way, edges represent an estimate of the total evolutionary 3D-cost-metric change between consecutive matching

instances. Note that the distance from the (0,0,0)=DNA-SIGNATURE vertex to the actual (x,y,z) coordinate of any matching instance represents a form of absolute evolutionary distance. Similarly, the length of an edge between consecutive nodes simply allows computing the magnitude of their internodal evolutionary distance by a straightforward application of the Phytogoras' theorem.

PHYLOGENETIC SORTING

Note that, in the pure sense, the following plot does NOT represent a PHYLOGENETIC TREE as edges between nodes (i.e., matching instances) relate differential cost-metrics among matching instances (as well as with respect to the DNA signature) as opposed to evolutionary relationships between matching instances. Note however, that the plot can be used as a STARTING POINT to derive a phylogenetic tree between the matching instances and the DNA signature.



BOT Distance 3D metric relationship between the DNA signature and unearthed matching instances. This plot can be used to derive a phylogenetic tree showing the evolutionary relationship between the DNA signature and matching instances.

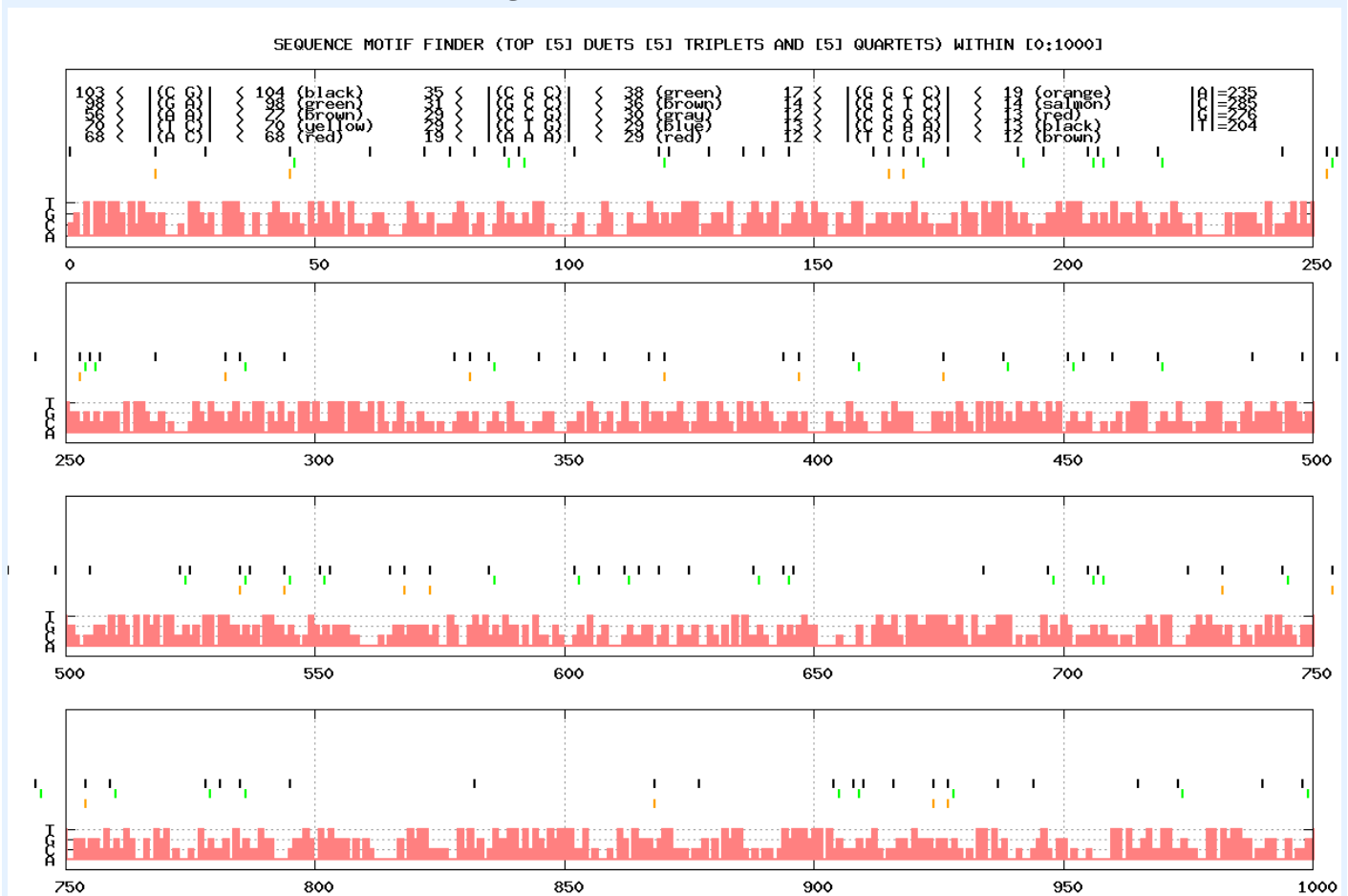
SEQUENCE MOTIF FINDER

"HPS DNA mining systems" provide additional functions such as the **sequence motif finder**. The sequence motif finder is a sequence-motif miner and visualizer that allows identifying the relative frequency and location for any, some, or even all DNA sequence-motifs DUETS, TRIPLETS, and/or QUARTETS (i.e., 2, 3, and 4-base). The sequence motif finder has linear run-time costs. For your convenience, "HPS DNA mining systems" applies the sequence-motif finder to the DNA signature by default - unless otherwise you specified a different subsequence interval.

SUMMARY INFORMATION FOR MOTIF DUETS, TRIPLETS, AND QUARTETS

The figure below shows the top (most frequent) sequence-motif DUETS, TRIPLETS, and QUARTETS found to lie within the specified subsequence of the DNA input sequence. As stated, the full extent of the DNA signature is examined by default, although it is possible to examine ANY subsequence of the **DNA input sequence**. Therefore, the subsequence interval analyzed is identified in the plot by a range notation of the form [#: #]. To allow closer examination, the selected interval of the DNA input sequence has been automatically divided into four subintervals of similar size being ordered (by increasing genomic address) from top to bottom. Summary data along the top part of the figure provides information about the relative frequency and identity of sequence-motifs found based on top-most frequency rank. Summary data is divided into four columns that corresponds (from left to right) to: (1) summary data for the top 5 motif DUETS, (2) summary data for the top 5 motif TRIPLETS, (3) summary data for the top 5 motif QUARTETS, and (4) summary data for the 4 DNA bases. By default, the top 5 sequence-motifs duets, triplets, and quartets are analyzed; however, this number can be specified. Moreover, for each sequence-motif, a **tight-bound interval** describes the number of **motif-repeats** found for said sequence-motif. In practicality, the lower-estimate (of the interval given) represents the true number of motif-repeats that corresponds to a sequence-motif. Furthermore, for each sequence-motif found subsequence, a color-code is assigned (note however that colors are re-used across columns). The motif's assigned color is used to color-code the display of **all repeats of its corresponding sequence-motif**. This way, each sequence-motif is shown by a stream of color-coded vertical arrows/lines placed in exactly **one row per sequence-motif**. Normally, all arrow-rows for sequence-motif DUETS are printed first, then all arrow-rows for TRIPLETS are printed, and finally, all arrow-rows for QUARTETS are printed. However, by default, the display of said motif-repeats arrow-rows is shown **only for the topmost sequence-motif** each for DUETS (top row), TRIPLETS (middle row), and QUARTETS (bottom row). The number of arrow-rows to print for the resultant DUETS, TRIPLETS, AND QUARTETS motif-set can be specified and its limited only by physical constraints of the plot.

SEQUENCE MOTIF FINDER



Repeats of the highest frequency sequence-motif(s) found to lie within the specified subsequence. Default setup displays summary (frequency and identify) data for only the TOP FIVE FREQUENCY sequence-motif DUETS, TRIPLETS, and QUARTETS (i.e., 2, 3, and 4 DNA-base sequences). The relative location for each of the motif-repeats of the TOPMOST DUET, TRIPLET, and QUARTET is also shown by means of color-coded arrow-rows. Both the number of (2, 3, 4-base) motif-repeat arrow-rows to display and the number of sequence-motifs to rank can be specified.

REGULAR-EXPRESSION MOTIF MINER

"HPS DNA mining systems" provide also the ability to search a given DNA sequence for instances of a matching subsequence with respect to a specified **motif-based regular-expression**. This regular expression applied against a sequence can be arbitrarily complex and when combined with the above motif finder, it is quite simple to construct. Complex sequence motifs can be specified with ease by non-specialists using a simple and easy to write **UNIX-like syntax/grammar** that allows the **specification of complex DNA patterns of repeated sequence-motifs of varying lengths**. The upcoming release of "HPS DNA mining systems" provides even more extensive regular-expression composition, discovery, and search tools.

SPECIFICATION OF MOTIF REGULAR EXPRESSIONS

A motif-based regular expression is specified as a **sequence of (one or more) tuple-pairs constraints** of the form "**(DNA_MOTIF)[NUMBER]**" where **(DNA_MOTIF)** represents a **(1, 2, 3, 4 DNA-base) sequence-motif specifier subsequence**. Valid examples of such are any of the following: (ACG), (CG), (AAAA), (ATA), (T), and (C). Note that the the motif must be delimited by opening and closing parenthesis. Similarly, **[NUMBER]** represents a positive number **specifier of the number of repeats** associated with the preceding sequence-motif. Valid examples of such are any of the following: [1], [3], [10], [100], and [*]. Note that the unknown number of repeats is specified by the special qualifier [*] and that numerical repeat qualifier must be delimited by opening and closing square brackets. This way, the following regular expression "**(AAT)[2](GG)[1](A)[3](C)[1](CGTA)[2]**" translates to a variable-length sequence-motif search against the specified sequence for the first subsequence of DNA-bases that FULFILLS ALL specified

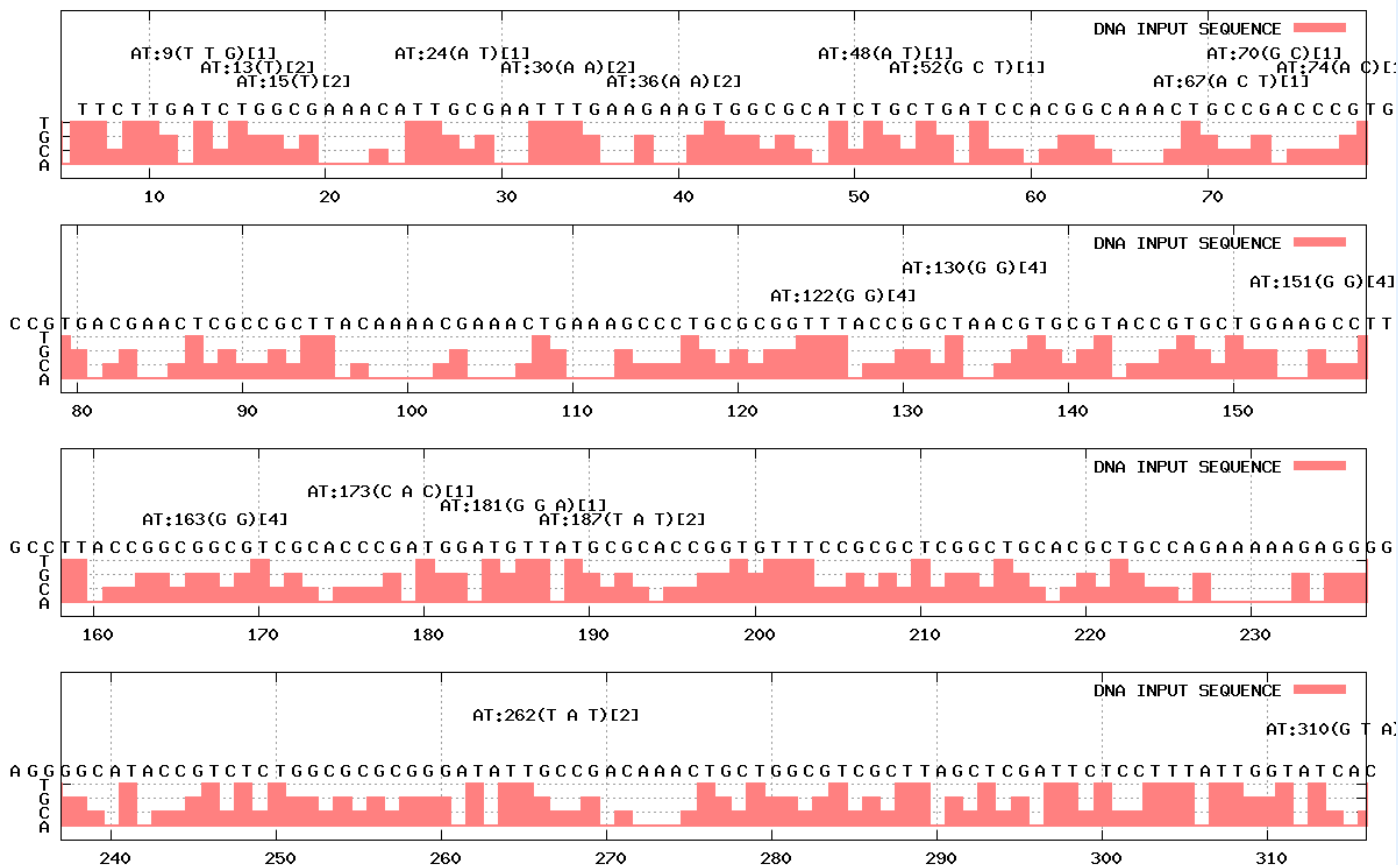
tuple-pair constraints found within the regular expression in the exact order given and with the exact number of motif-repeats specified. For example, assume the symbol "*" represents a random DNA base, then the above specified motif-based regular expression would be matched against a subsequence such as the following: "AAT***AAT*****GG***A*****A*A***C***CGTA***CGTA" if one such like exists from within the specified interval of the DNA input sequence. The motif-based regular expression finder has linear-time run-time complexity.

SUMMARY INFORMATION FOR MOTIF REGULAR EXPRESSION MINING

The following figure depicts the results of a regular expression search within the DNA input sequence. By default, the motif-based regular expression is applied to (the full extent of) the DNA signature found within the DNA input sequence. However, any given interval-based subsequence of the DNA input sequence can be examined by the motif-based regular expression miner. Moreover, one or more matching instance can be extracted (if any) but by default, only the very first match from within the specified interval is shown. The plot automatically zooms in into the MAXIMUM DETAIL POSSIBLE for the interval of the matching subsequence found to match all the tuple-pair constraint of the motif-based regular expression. Moreover, said resultant matching interval is further divided into FOUR EQUAL LENGTH CONTIGUOUS SUBINTERVALS, which are then plotted one per panel, with the first of said subintervals shown at the top and correspondingly, the last subinterval at the bottom. The motif-based regular expression that was searched-for is printed atop of the figure while the precise location of each (thus) successfully met tuple-pair constraint of the motif-based regular expression is shown within the exact location of the matching subinterval plot. Each DNA base of the matching subinterval is printed (along a single row across all subplots) and above the relevant location of a precise match of a tuple-pair constraint in a subplot, the actual tuple-pair constraint being met is printed (across any of three allotted tuple-pair constraint rows). Note that for tuple-pair constraints specified a repeat, the tuple-pair constraint is shown the corresponding that many times on the relevant subinterval and subplot. Note the precise and exacting correspondance - as with any of the features of our system.

DNA MINING BASED ON MOTIF REGULAR EXPRESSIONS

REGULAR-EXPRESSION: (TTG)[1](T)[2](AT)[1](AA)[2](AT)[1](GCT)[1](ACT)[1](GC)[1](AC)[1](GG)[4](CAC)[1](GGA)[1](TAT)[2](GTA)[1]



Resultant matching subsequence interval associated with the successful mining of the specified motif-based regular expression against the specified interval of the DNA input sequence. The motif-based regular expression miner handles motif-repeats of variable-length across variable gap lengths. Any regular expression consisting of (1, 2, 3, 4-base) sequence-motifs can be specified. Regular expressions follow an easy Unix-like model described above.

ABOUT "HPS DNA MINING SYSTEMS"

"HPS DNA mining systems" are implemented in COMMON LISP with 100% proprietary in-house code that has been designed for maintenance and extensibility. Moreover, "HPS DNA mining systems" is not just an application but rather an easily extensible programming environment for bioinformatic applications. The HPS code base has been designed with portability in mind, with neither operating system nor compiler dependencies. Use of external (non-incorporated into) and unmodified applications GNUplot (*Copyright 1986 - 1993, 1998, 2004 Thomas Williams, Colin Kelley*) and GNUwget (*Copyright © 1996-2005 Free Software Foundation, Inc.*) is made under respective licenses ([GNUPLOT license](#)) and ([GNUWGET license](#).)

HPCC SUITABILITY

Current implementation has been tested on the Windows XP operating system under typical Pentium-class personal computer power. Even though "HPS DNA mining systems" are specifically designed to allow data mining of large DNA databases on Pentium-class personal computers, "HPS DNA mining systems" are inherently suitable for future deployment on HPCC's grid-computing systems through batched pre-computation and adaptive parceling of key "HPS DNA mining systems" tasks.

DISCLAIMER

This report was automatically generated by "HPS DNA mining systems". "HPS DNA mining systems" are based on proprietary patent-pending technologies developed by Dr. Nelson R. Manohar-Alers. No permit to use "HPS DNA mining systems" algorithms, software, technologies, disclosures, or derivatives of such is granted without authorization from Dr. Nelson R. Manohar. The research direction underlying "HPS DNA mining systems" are a result of the unique blend, refinement, and reuse of skills (accumulated by Dr. Manohar-Alers across 20 years of professional experience) to continuously achieve an innovative leaping furtherance of our initial (c. 1992) long-term-held research thread on adaptive systems. **Qualified principals and/or reviewers from qualified institutions** may present inquiries by e-mailing the author (**Dr. Nelson R. Manohar, Principal Research Scientist, and Principal of "HPS DNA mining systems"**) at the e-mail address given atop this report. Dr. Nelson R. Manohar is open to consider "qualified and appropriate" collaboration, seeding, funding, or contract inquiries from **qualified principals** from **reputable and qualified (US/EU) institutions**. However, because of the nature, volume, intensity, and disclosure-level of our (environments and) research work, we reserve the right to reply to any inquiry. For example, inquiries that somehow could be related to disclosure of patent-pending "HPS DNA mining systems" may be politely ignored. Moreover, inquiries from corporate laboratories may not be answered.

HPS research work has entirely been self-funded and achieved under sustained and adverse malfeasance conditions. "HPS DNA mining systems" research work did NOT received support from either the **National Science Foundation (NSF)**, the **National Institutes of Health (NIH)**, or any other similar granting agencies. It is relevantly disclosed that **(U.S.) FEDERAL AND (N.Y., P.R.) STATE MALFEASANCES** have previously been documented, disclosed, and filed and such relate to matters of STRONG NATIONAL INTEREST. Please inform the Inspector General of the U.S. Department of Justice (Mr. Glenn A. Fine) (at askdoj@usdoj.gov) of any "suspected-to-be malfeasance order" interfering with inquiries to us on areas related to ("but not limited to") research, funding, employment, collaborations, etc. Finally, we will politely ignore inquiries from parties suspected to have compromised, malfeasance, or political interests -- proper outlets for such have been given. Dr. Nelson R. Manohar-Alers is a U.S. Citizen.

This webpage is best seen WITHOUT font/color substitution at **1200x800** pixel resolution (e.g., the WIDESCREEN/LANDSCAPE format typically used in laptop displays). This webpage is also designed to fit and print on PORTRAIT format as long as font override is NOT specified. Therefore, for best viewing and printing, you should have the **"Verdana"** font installed in your system. Plots found within use the PNG (Portable Graphics) format and are generated at the relatively high (1200x800) pixel resolution. OutThis resolution allows for reasonable magnification detail in most image editors or modern web browsers. Some links on our

reports are set to (by default) open as a new browser window or browser tab; therefore, check that your computer does not mistake such opening page as an apparent popup window, it is not. Our webpages and sites contain NEITHER scripts NOR active code. Display of this webpage has been tested for the Opera browser on a Windows XP platform. We value your copyrights, please respect ours.

(c) Nelson R. Manohar, Ph.D. - 2005-2007. All rights are reserved.