THE HPS DNA (HPS-SHAPE-APPROXIMATE) SEQUENCE ALIGNMENT AND PATTERN DISCOVERY SYSTEM

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

REPORT OF HPS DNA SEQUENCE ALIGNMENT

The functions on this file implement an approximate pattern match search for a signature within test sequence. Both the signature and test sequence are time series of HPS_DNA_ segments. Both these were generated using the HPS_DNA_ transform at common baselining values (e.g., HPS_DNA_(60, 30, 0.001)). The approximate pattern match search looks for all the possible instances of the signature within the test sequence. A match is produced when a significant number of ordered HPS_DNA_ segments in a signature are approximately matched to an appropiate set of HPS_DNA_ segments in the test sequence. The match is a (conditional) probabilisty match based on the combined probability of an ordered set of HPS_DNA_ segments in the signature to approximately match an ordered set of HPS_DNA_ segments in the test sequence. To do this, each individual approximate match attempts to find for each segment in the signature, an instance within the test sequence which has similar duration, relative positioning within the time series, and similar HPS_DNA_ targeting value. All repeated approximate repeats of the presence of the signature within the test sequence are found and reported. Note that even overlapping instances are also uncovered. The overall complexity of the approximate pattern match is as follows:

if n is the number of HPS_DNA_ segments in the signature and m is the number of HPS_DNA_ segments in the test sequence the algorithm has O(m(m*n)) time complexity (i.e., time complexity is quadratic on the HPS_DNA_ fractality of the test sequence). This relatively low complexity is achieved through multiple space tradeoffs. The reader should note that this complexity is relatively low once one realizes that m and n are HPS_DNA_ fractalities, and thus could often be several orders of magnitude in reduction from the size of their underlying time series. For example, for a test sequence such as the DNA which prior to HPS_ DNA_ encoding was of size equal to 1100 units, its HPS_DNA_ approximation had only 20 or so non-trivial HPS_DNA_ segments. If the test sequence had 60 non-trivial segments, then algorithm complexity is of the order of O(60 * 60 * 20) which compares favorably to O(3600^2) complexity needed to deterministically find approximate substrings. This code is for demonstration purposes and is not intended as an efficient implementation of the approximate pattern match but rather as an illustration of such being possible, something which was argued could not be done by those who like to critique hoping to sabotage credibility.

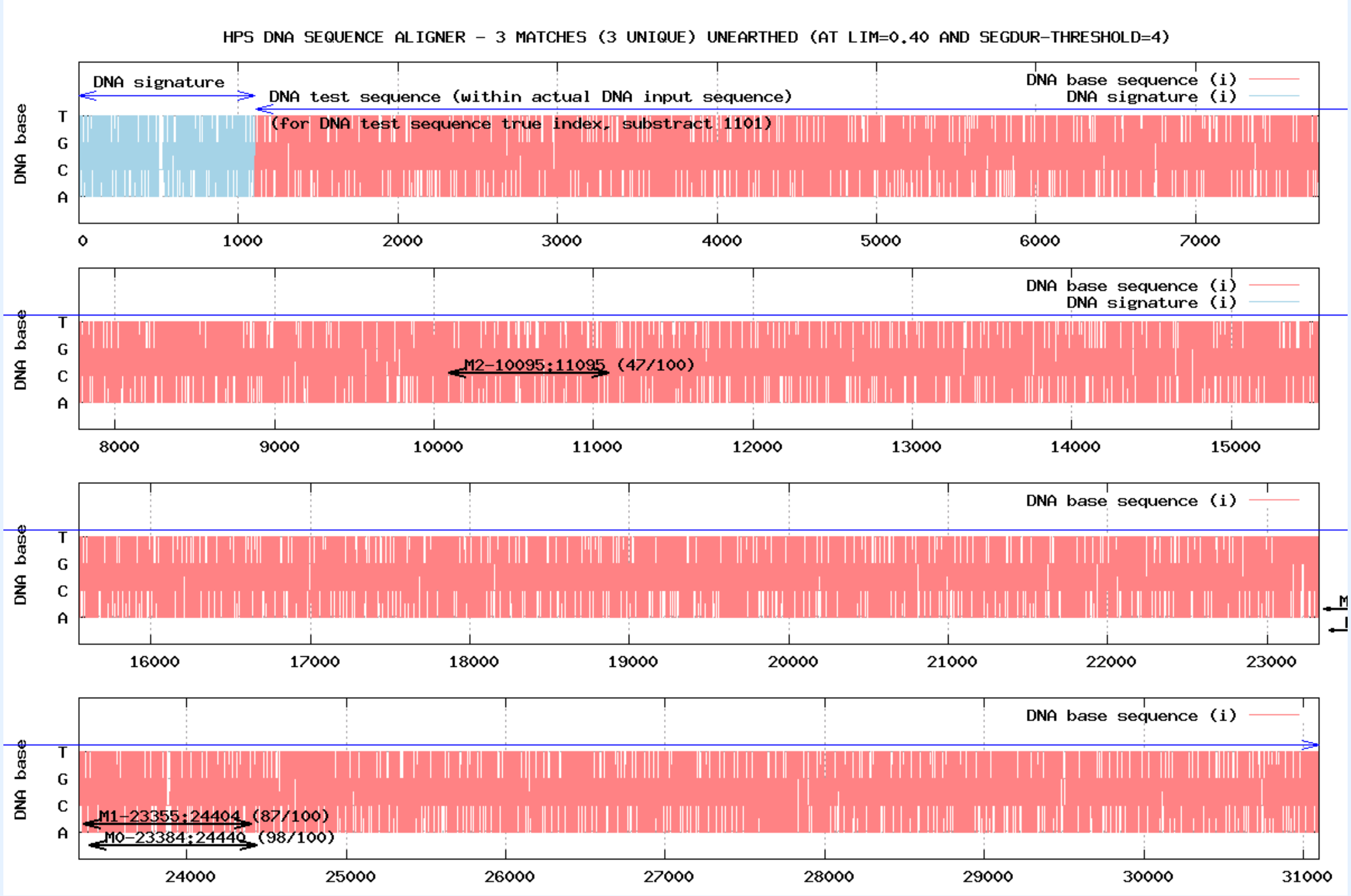
Each instance is ranked in terms of a cost metric which provides a relative comparison of how good an approximate pattern match is an individual instance. The user is provided the opportunity to reduce the output by choosing a threshold point, for which approximations exhibiting a cost metric higher than such are discarded. The output consist of a table of the resulting approximate pattern match for repeats of the HPS_DNA_ signature within the test sequence together with cost metrics information. Of course, the (conditional) probability of the approximate pattern match could also be computed but that is not done yet.

The approximation produced by the HPS TRANSFORM results in a time series of "ATS segments" that exhibit small and controlled variability around the (hidden) fundamental frequencies of its underlying random process. That is, without explicit awareness of TIMESCALE on the input signal, the time series of these random-length "ATS segments" describes a bounded-error controlled variability trajectory over the (true but hidden) "PROCESS STATES" of the underlying random process {X}. Below is shown a plot of the resultant HPS approximation.

BEHAVIOR OF THE PATTERN MINING PROCESS ACROSS TIME

The approximation produced by the HPS TRANSFORM results in a time series of ATS segments that exhibit small and controlled variability around the (hidden) fundamental frequencies of its underlying random process. That is, without explicit awareness of TIMESCALE on the input signal, the time series of these random-length ATS segments describes a bounded-error controlled variability trajectory over the (true but hidden) PROCESS STATES of the underlying random process {X}. Below is shown a plot of the resultant HPS approximation.

BEHAVIOR OF THE OVERALL PATTERN MINING PROCESS - Figure 2



TOP

Time plot of the resultant HPS match strength along the entire traversal of the input timeseries. The presence of a true match is observed as a heavy-tail event among a noise background.

2ND

Time plot of the observed HPS segment duration for the resultant HPS DNA approximation, where longer spikes represent longer HPS ATS segments

3RD

Time plot of the resultant HPS approximation to the input signal. The relative time index of the input signal (DNA sequence) is preserved. As a result, the first M elements represent the HPS signature and the remainder N-M represent the HPS test sequence.

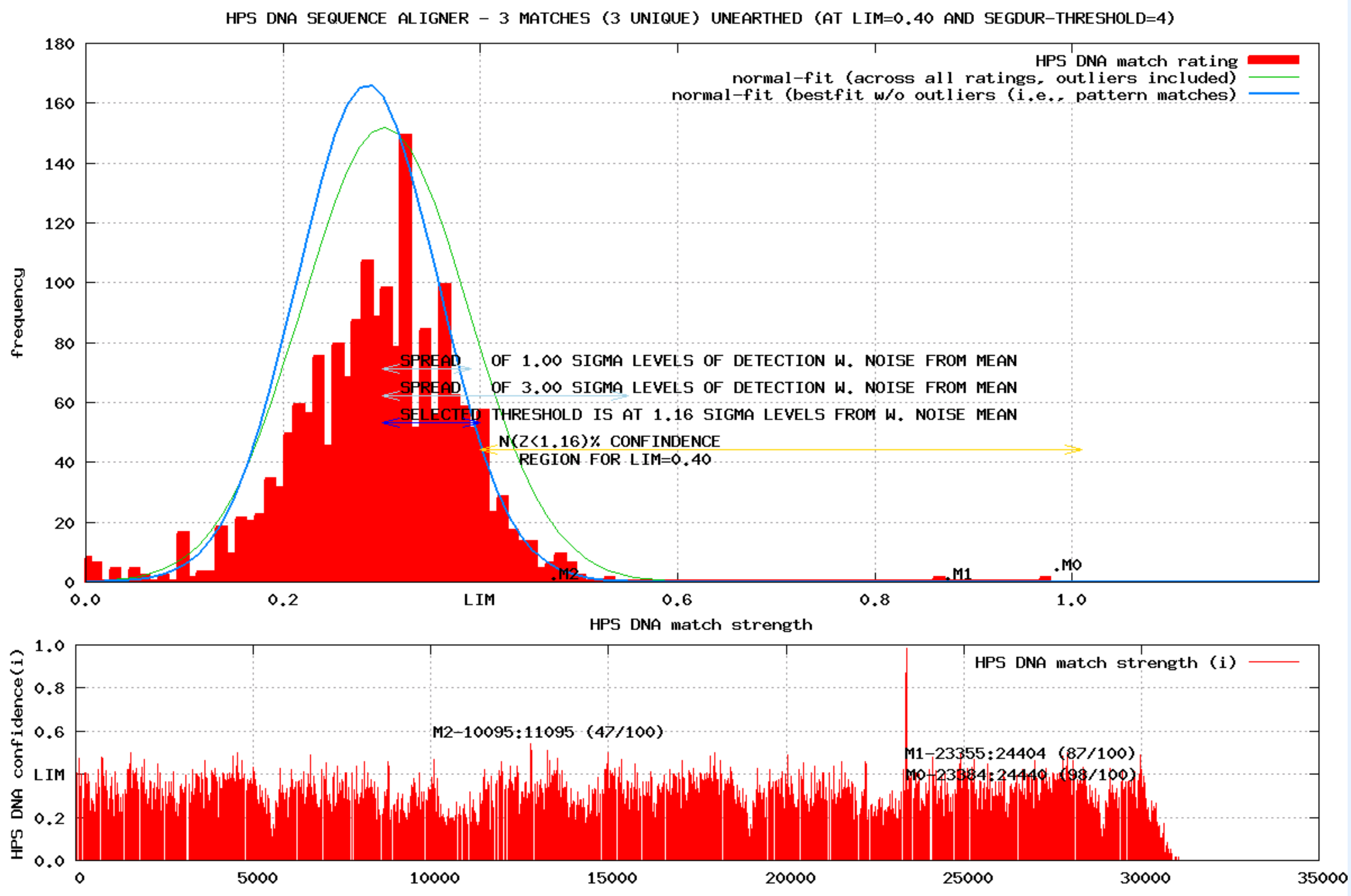
BOT

Time plot of the input signal - that is, the input DNA sequence - in our current HPS mining format. In this format, the input signal is constructed as follows: the first M elements represent the DNA signature to data mine and the remainder N-M elements represent the DNA test sequence to be mined for the presence of the DNA signature.

RELATIVE STRENGTH OF THE RESULTANT PATTERN MATCHES

Note how the unearthed pattern matches lie as heavy tail outliers with respect to the rest of the match rating signal values. Match rating values for non-matches bunch as gaussian white noise in the lower range whereas match rating values for true matches lie well above a 3 SIGMA level limiting value (pointed to by LIM). The algorithm behaves as a white noise filter remover for non-matches while for true matches detects then with high SNR values (well above the predominant noise levels).

HISTOGRAM OF MATCH RATING SIGNAL VALUES ACROSS ALL PIVOT POINTS IN THE DNA TEST SEQUENCE



GREEN-FIT
Resultant normal curve for mean and sigma parameters derived from the data (when heavy tail outliers are included).

TOP
Best normal curve fit for resultant mean and sigma parameters (when heavy tail ooutliers are removed)points.

M0: DETAILED REPORT FOR APPROXIMATE PATTERN MATCH

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	COST_METRIC	COST_VECTOR
23384	24440	0.98214287	1057	877.0	0	(2 1 51 0 0 1 0 0)

E1	TOTAL NUM: 2	(EXACTING_MATCH)	EXACTING SEG-DUR, SAME SEG-VAL MATCHES
E2	TOTAL NUM: 1	(APPROX32_MATCH)	APPROX. ~SEG-DUR, SAME SEG-VAL MATCHES
E3	TOTAL NUM: 51	(APPROX64_MATCH)	LOOSELY ~SEG-DUR, SAME SEG-VAL MATCHES
E4	TOTAL NUM: 0	(SPANNING_MATCH)	SPANNING SEG-DUR, SAME SEG-VAL MATCHES
A1	TOTAL NUM: 0	(EXACTING_MISPL)	EXACTING SEG-DUR, +-1MISPELLED SEG-VAL
A2	TOTAL NUM: 1	(APPROX32_MISPL)	APPROX. ~SEG-DUR, +-1MISPELLED SEG-VAL
A3	TOTAL NUM: 0	(APPROX64_MISPL)	LOOSELY ~SEG-DUR, +-1MISPELLED SEG-VAL

A4	TOTAL NUM: 0	(SPANNING_MISPL)	SPANNING SEG-DUR, +-1MISPELLED SEG-VAL

M1: DETAILED REPORT FOR APPROXIMATE PATTERN MATCH

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	COST_METRIC	COST_VECTOR
23355	24404	0.86885244	1050	1019.0	0	(2 0 49 0 0 0 2 0)

E1	TOTAL NUM: 2	(EXACTING_MATCH)	EXACTING SEG-DUR, SAME SEG-VAL MATCHES
E2	TOTAL NUM: 0	(APPROX32_MATCH)	APPROX. ~SEG-DUR, SAME SEG-VAL MATCHES
E3	TOTAL NUM: 49	(APPROX64_MATCH)	LOOSELY ~SEG-DUR, SAME SEG-VAL MATCHES
E4	TOTAL NUM: 0	(SPANNING_MATCH)	SPANNING SEG-DUR, SAME SEG-VAL MATCHES
A1	TOTAL NUM: 0	(EXACTING_MISPL)	EXACTING SEG-DUR, +-1MISPELLED SEG-VAL
A2	TOTAL NUM: 0	(APPROX32_MISPL)	APPROX. ~SEG-DUR, +-1MISPELLED SEG-VAL
A3	TOTAL NUM: 2	(APPROX64_MISPL)	LOOSELY ~SEG-DUR, +-1MISPELLED SEG-VAL
A4	TOTAL NUM: 0	(SPANNING_MISPL)	SPANNING SEG-DUR, +-1MISPELLED SEG-VAL

M2: DETAILED REPORT FOR APPROXIMATE PATTERN MATCH

MATCH_STARTS	AND_ENDS_AT	MATCH_RATING	MATCH_SPANS	MATCH_WEIGHT	COST_METRIC	COST_VECTOR
10095	11095	0.47058824	1001	1883.0	0	(2 7 5 1 1 1 7 0)

E1	TOTAL NUM: 2	(EXACTING_MATCH)	EXACTING SEG-DUR, SAME SEG-VAL MATCHES
E2	TOTAL NUM: 7	(APPROX32_MATCH)	APPROX. ~SEG-DUR, SAME SEG-VAL MATCHES
E3	TOTAL NUM: 5	(APPROX64_MATCH)	LOOSELY ~SEG-DUR, SAME SEG-VAL MATCHES
E4	TOTAL NUM: 1	(SPANNING_MATCH)	SPANNING SEG-DUR, SAME SEG-VAL MATCHES
A1	TOTAL NUM: 1	(EXACTING_MISPL)	EXACTING SEG-DUR, +-1MISPELLED SEG-VAL
A2	TOTAL NUM: 1	(APPROX32_MISPL)	APPROX. ~SEG-DUR, +-1MISPELLED SEG-VAL
A3	TOTAL NUM: 7	(APPROX64_MISPL)	LOOSELY ~SEG-DUR, +-1MISPELLED SEG-VAL
A4	TOTAL NUM: 0	(SPANNING_MISPL)	SPANNING SEG-DUR, +-1MISPELLED SEG-VAL

ABOUT THE HPS TRANSFORM

The **HPS Transform** allows the unearthing of a form of timescale information from a signal. Although the **HPS Transform** is particularly suited for adaptive process control, the results have vast implications to other fields. In particular, important applications of the **HPS Transform** relate to feature extraction and data-mining. For example, the **HPS Transform** is particularly useful in distributed, loosely-coupled, adaptive process control applications where a client *C* performs monitoring of a local resource *R* with sampling effort □□*y(i)*□□ but now needs to report to a remote server *S only* only on certain changes over a state memory □*π(k)*□*,* thus exhibiting a reduction property □*k*□ *significantly less than* □*i*□. Intuition into the operation region of the **HPS Transform**. The **HPS Transform** exhibits most desirable qualities on implementation ease, algorithmic complexity (i.e., O(N) - linear time worst time computational time), computational stability (i.e., stable operation region), signal compressibility, (e.g., large ratio, controllable error/compression ratio), decision-making robustness (e.g., robustness to outliers, timescale independent, robustness to lack of information about the input signal's distribution, etc.) information loss, and error behavior (i.e., bounded error and known confidence levels). The HPS transform achieves this by trading off a small but specifiable delay, this being measured in terms of samples (and not time).

DISCLAIMER

This report is automatically generated by the HPS DNA pattern miner, based on the HPS Transform, and the HPS stenographic decoder. No permit to use is granted without authorization from the author. Due to malfeasance, the stenographic application can not be made available for anonymous use. Only qualified reviewers and academics may request access to the application by e-mailing your specific request to the author to the e-mail address given.