Advancements in scheduling simulated annealing for network alignment

Hudson Hughes*

Dr. Wayne Hayes, Bren School of Computer Science, University of California, Irvine

E-mail: hlhughes@uci.edu

Abstract

Several attempts have been made to tackle the problem of network alignment, which has applications in social network analysis, bioinformatics, and more. Most methods leave much to be desired in terms of memory usage, execution time, and the heuristic value of the final alignment. SANA (Simulated Annealing Network Alignment) is a new probabilistic search algorithm designed to address the problem and it seems to be superior to every alternative. An important aspect of simulated annealing is the creation of a temperature schedule, which controls the likeliness of traversing poor solutions during the search for the optimal one. Finding an ideal way to create schedules has been an iterative process. The objective of this paper is to demonstrate the effectiveness of a new method that is quickly and reliably able to provide productive schedules using binary search and linear regression.

Introduction

An alignment refers to a node correspondence between two separate networks. Network alignment is the process of searching for the correspondence with the highest score, which is based on a heuristic that changes from application to application. As one might expect, the

number of alignments increases exponentially with the size of the networks, so checking the score of every possible alignment is impractical. A more efficient search method is necessary.

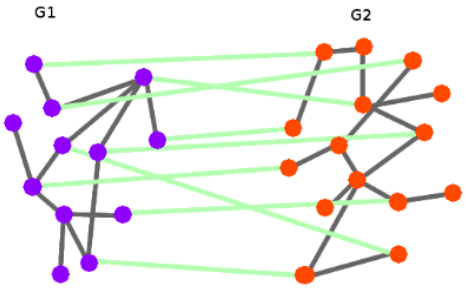

Figure 1: Visual Representation of Network Alignment

For each alignment, a neighboring alignment can be found by swapping one of the nodes in the correspondence. All of these connected neighbors form a "landscape" of solutions where their quality is represented by height. Hill climbing and simulated annealing are methods that use this knowledge. The former is the process of randomly selecting a solution, then repeatedly replacing the selected solution with a superior neighbor until a "peak" is reached or time expires. Simulated annealing is an evolution of hill climbing as its design acknowledges that a peak found with hill climbing might not be the best solution in the landscape. The iterative optimization technique introduces the idea of probabilistically selecting non-optimal solutions in order to allow the algorithm to search more of the landscape. SANA's checking and traversing of each alignment will be referred to as an iteration.

In SANA, the likelihood of accepting a worse solution is based on the difference in quality between the two alignments and a floating point variable that is referred to as the temperature. Larger differences and lower temperatures create smaller probabilities of accepting poor solution and vice versa. As SANA progresses, the temperature decreases from an initial temperature called t-Initial and a final temperature called t-Final. For each network pair and heuristic combination, there is an optimal pair of temperatures to move between. Creating a temperature schedule means to find those temperatures. Users can specificy a

schedule manually, but the amount of guesswork and mathematics required necessitate that the process be automated.

Given a proper temperature schedule, a run of the SANA program will begin with a near 1.0 probability of accepting alignments that are worse than whatever is currently selected. By the end of the run, that probability will have gradually dropped down to approximately 0. The probability that a temperature yields is referred to its pBad. A temperature's pBad can only be estimated by performing a brief SANA run with the temperature and empirically meausre the pBad. These three algorithms that use pBad approximations to create productive temperature schedules are compared.

Previous Work

Linear Regression Method

The previous method for creating a temperature schedule used linear regression on the relationship between temperature (which will be represented on the X-axis) and pBad (which will be represented on the Y-axis). This is called the TP (temperature-pBad) relationship and it is typically viewed in logarithmic space. The temperature space between 1E-10 and 1E10 is searched on the assumption that the range is wide enough to capture the temperatures required. Also, it can safely be assumed that for each valid network pair, the TP relationship follows a trend where the pBad starts very low at low temperatures before jumping to nearly 1.0. By taking 100 logarithmically and evenly spaced samples between 1E-10 and 1E10 from a TP relationship, the regression can be used to calculate the three lines that summarize the entire relationship. The right end of the center line is used as a starting place to search for a t-Initial and that temperature is increased exponentially until the corresponding pBad is at least 0.985. Likewise, the left end of the center line is used as a starting place to search for a t-Final and that temperature is decreased until the corresponding pBad is under 1E-6. The selection of these target probabilities are semi-arbitrary. The idea is that if the starting

temperature is too high, SANA will waste a lot of time tolerating poor alignments. If the t-Final is too low, the simulated annealing will turn into an instance of hill climbing too quickly. Spending too much time hill climbing becomes pointless as it saturates quickly.

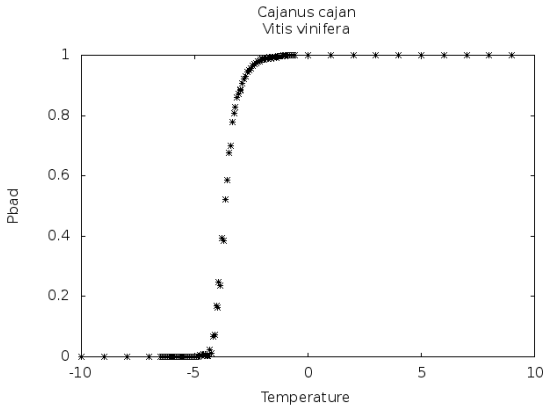

Figure 2: 100 samples of a TP relationship

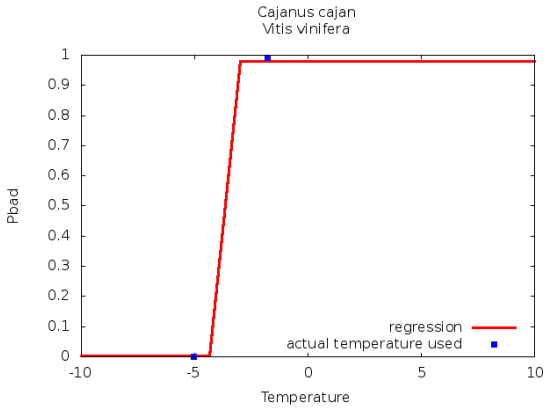

Figure 3: Linear regression of the same TP relationship

While this algorithm succeeds in creating productive temperature schedules, it requires at least 100 temperatures to be tested for their corresponding pBad, each of which takes at least one second. The default amount of time SANA is given to search for an alignment is five minutes and some users will want results in an even smaller amount of time. When this is taken into consideration, an initialization step that is nearly two minutes at best becomes problematic. Still, the linear regression method is a reliable method to compare faster alternatives with. Also, the data provided using the algorithm illustrates a relationship that is important to understanding how SANA is expected to progress.

Binary Search Method

The next method for creating a temperature schedule builds on the assumptions made during linear regression. The TP relationship is still analyzed, but many of the calculations are skipped in comparison to the previous algorithm. It starts by initializing a temperature variable called index at 1E-10. It is then increased by a factor of 10 until a corresponding pBad greater than 1E-6 is encountered, this process is known as a linear search. At that

point, the current and previous value of index are used as bounds for a binary search to find a temperature that yields a pBad between 1E-6 and 1E-7. This temperature will be used as t-Final.

Listing 1: binary search algorithm

```
binarySearchLeftEnd = index - 1
binarySearchRightEnd = index
mid = (binarySearchRightEnd + binarySearchLeftEnd) / 2
while 10^binarySearchRightEnd / 10^binarySearchLeftEnd > 2:
    temperature = 10^mid
    initialProbability = pbadForTemp(temperature)
    if initialProbability > 1E-6:
        binarySearchRightEnd = mid
        mid = (binarySearchRightEnd + binarySearchLeftEnd) / 2
    else if initialProbability < 1E-7:
        binarySearchLeftEnd = mid
        mid = (binarySearchRightEnd + binarySearchLeftEnd) / 2
    else:
        break
startingExponent = (10^binarySearchRightEnd / 10^binarySearchLeftEnd) < 2 ? binarySearchRightEnd : mid
TInitial = 10^startingExponent
```

The same process is used to get t-Initial, but 1E-7 is replaced by 0.985 and 1E-6 is replaced by 0.995.

The binary and linear search combination is expected be able to find useful temperatures similar to that of linear regression with significantly fewer pBad test, and thus, significantly less time.

Even with the reasonable expectation that fewer pBad measurements will be needed to create a schedule, the exact amount of measurements could still vary widely between runs as the search is still probabilistic. It'd be best to have a fixed amount so that users can easily predict how long the algorithm will take to complete.

Combination Method

This new temperature schedule algorithm involves performing the same linear and binary search as previously detailed; except the index variable iterates all the way through 1E10 and all of their corresponding pBads are sampled. Also, the binary searches are limited to

taking 8 pBad test before exiting. This way there is a predictable limit of how many tests are taken. If the binary search is cut off too early then there won't enough samples. If the search is allowed to go on for too long then there will be an excess of samples that close in on a single point. Next, a 3-line linear regression of every TP relationship sample taken is made and assert that the slope of the middle line is greater than zero; meaning, if it is not the program will terminate. The reasoning behind this is that in productive simulated annealing the tolerance for poor solutions must continually decrease with the temperature. It can be concluded that this wont be the case if the TP relationship is shaped like a backwards "Z". Finally, the list of samples is analyzed and the lowest temperature with a pBad above 0.985 is assigned to be the t-Initial and the highest temperature with a pBad below 1E-6 assigned to be the t-Final. This algorithm is guaranteed to finish after 38 samples are taken from the TP relationship and it can be proven that it can quickly provide optimal temperature schedules while filtering invalid network pairs.

Temperature decay mechanics

SANA follows a specific pattern to traverse between temperature boundaries after they have been set. First, the speed at which SANA is able to traverse solutions is approximated and this data is used to predict how many iterations SANA will go through in the allotted time. A variable called tDecay is declared as a function of the temperature boundaries. Given the amount of iterations SANA may be completed at a given time and the previously mentioned numbers, the corresponding temperatures can be calculated.

$$TDecay = -log(t - Final/(t - Initial))$$

Figure 4: TDecay equation

As a result, the temperature decreases in a sigmoid function as SANA runs. As long as a SANA run's speed remains constant the temperature will equal t-Final once time expires.

$$T = exp(-TDecay * currentIteration/expectedTotalIterations)$$

Figure 5: Temperature equation

Methods

There are three established methods to create simulated annealing temperature schedules: one using linear regression, another that uses binary searches, and a third that is a combination of the first two. The next step is to determine how well they perform in terms of execution time and reliability. A method's reliability is its ability to continuously supply temperatures that correspond to pBads close to the ideal ones that was previously defined (0.985 and 1E-6). Eight networks were taken from the Biogrid dataset and 28 network pairs were created from them. Also, 500 larger networks were curated from the Toronto data set and 99 more pairs were made. SANA is then run with these 127 pairs using the three previously mentioned scheduling methods and the S3 heuristic function for a total of 381 runs. Several variables from the progress of the runs are saved and placed in tables or charted including:

- the amount of time it took to calculate the schedule
- the pBad samples and linear regressions so that plots of the second and third algorithm can be compared to that of the first
- the t-Final, t-Initial, their pBads, and the score of the final alignment

The expectation is that the alignments SANA generates using each of the temperature schedule algorithms will be of similar quality. The main difference should come from the execution time of each scheduling algorithm. A confirmation that linear regressions created with the new method are similar to that of the old method is also expected.

Results

Across all 127 network pairs, SANA has performed equally well using all of the temperature scheduling methods. In logarithmic space, the temperatures each method settles on are reasonably close. The corresponding pBads of these temperatures are also ideal. The pBads of the initial temperatures are above 0.985 and that of the final temperatures are usually below 1E-6.

Table 1: Final alignment score comparison across Biogrid pairs

	Combination	Binary Search	Regression
AThaliana CElegans	0.367381	0.385092	0.381734
AThaliana DMelanogaster	0.269179	0.262357	0.267008
AThaliana HSapiens	0.30522	0.283036	0.294836
AThaliana MMusculus	0.310413	0.332044	0.31289
AThaliana RNorvegicus	0.586151	0.588531	0.572135
AThaliana SCerevisiae	0.106922	0.100765	0.104521
AThaliana SPombe	0.480919	0.451267	0.493068
CElegans DMelanogaster	0.367517	0.372134	0.354984
CElegans HSapiens	0.42517	0.411049	0.416667
CElegans MMusculus	0.359705	0.357542	0.343499
CElegans RNorvegicus	0.415633	0.424066	0.421842
CElegans SCerevisiae	0.305931	0.314463	0.317659
CElegans SPombe	0.360207	0.356719	0.360935
DMelanogaster HSapiens	0.193645	0.192641	0.194737
DMelanogaster MMusculus	0.301435	0.301859	0.288859
DMelanogaster RNorvegicus	0.474092	0.41794	0.489336
DMelanogaster SCerevisiae	0.13192	0.130664	0.132591
DMelanogaster SPombe	0.434326	0.422276	0.429967
HSapiens MMusculus	0.345143	0.34507	0.356251
HSapiens RNorvegicus	0.675566	0.663039	0.735381
HSapiens SCerevisiae	0.173191	0.177635	0.172463
HSapiens SPombe	0.502681	0.49246	0.492079
MMusculus RNorvegicus	0.560302	0.563156	0.518797
MMusculus SCerevisiae	0.257774	0.257183	0.259131
MMusculus SPombe	0.43087	0.437947	0.436901
RNorvegicus SCerevisiae	0.685238	0.679032	0.667074
RNorvegicus SPombe	0.446704	0.461831	0.456963
SCerevisiae SPombe	0.453559	0.456143	0.453198
Average	0.3831	0.379926	0.383054

In terms of execution time and resources, the binary search method proved to be superior to the others, which was unexpected. The criteria of the the search was designed to keep it from running indefinitely, which is a danger when performing probabilistic searches. It also keeps the program from performing any more than 17 pBad tests. The amount of time required to complete a test is a function of the networks' sizes, but that time is uniform across each method. The new combination method always requires 29 tests, and the old linear regression method requires 121 on average. This information translates to the actual execution time. The binary search method and combination method are huge improvements over the original linear regression method.

Even after a temperature schedule is established, it is not guaranteed that SANA will

Table 2: Initial temperatures and pBads calculated using each method presented in logarithmic space

	Combination	Binary Search	Regression	Regression After Adjustment
AThaliana CElegans	-3.5 0.990185	-3.5 0.990711	-4 0.992444	-3.6 0.990794
AThaliana DMelanogaster	-3.75 0.990954	-3.75 0.99059	-4.425 0.992346	-3.725 0.991527
AThaliana HSapiens	-4 0.987853	-3.75 0.993834	-4.35 0.994528	-3.95 0.991325
AThaliana MMusculus	-3.5 0.990179	-3.5 0.994444	-4.05 0.991772	-3.55 0.990929
AThaliana RNorvegicus	-2.875 0.992814	-2.75 0.990458	-3.35 0.993926	-2.85 0.993814
AThaliana SCerevisiae	-3.125 0.98689	-2.75 0.994378	-4 0.987061	-3 0.990273
AThaliana SPombe	-3.125 0.995012	-3.25 0.993139	-3.6 0.995703	-3.4 0.99318
CElegans DMelanogaster	-3.75 0.985269	-3.75 0.98847	-4.125 0.992918	-3.625 0.992129
CElegans HSapiens	-3.625 0.9884	-3.5 0.996186	-4.05 0.991288	-3.55 0.991573
CElegans MMusculus	-3.25 0.994643	-3.5 0.987061	-3.975 0.987948	-3.275 0.994301
CElegans RNorvegicus	-2.875 0.99253	-2.5 0.998049	-3.4 0.99129	-2.7 0.993969
CElegans SCerevisiae	-3.25 0.994458	-3.5 0.988248	-4.125 0.985031	-3.425 0.991986
CElegans SPombe	-3.125 0.990389	-3.25 0.989489	-3.5 0.996415	-3.2 0.990275
DMelanogaster HSapiens	-4.125 0.986445	-3.75 0.995354	-4.5 0.995554	-4.1 0.99021
DMelanogaster MMusculus	-3.75 0.992413	-3.75 0.989941	-4.2 0.99484	-3.8 0.993624
DMelanogaster RNorvegicus	-3.125 0.993607	-3.25 0.991338	-3.6 0.993747	-3 0.996464
DMelanogaster SCerevisiae	-3.5 0.992191	-3.75 0.987622	-4.3 0.993281	-3.6 0.993037
DMelanogaster SPombe	-3.5 0.988937	-3.5 0.988685	-3.75 0.995094	-3.35 0.994615
HSapiens MMusculus	-3.75 0.990788	-3.5 0.994887	-4.125 0.993297	-3.825 0.990045
HSapiens RNorvegicus	-3.0625 0.991575	-2.75 0.992026	-3.3 0.992485	-3.1 0.992029
HSapiens SCerevisiae	-3.75 0.988454	-3.75 0.989118	-4.2 0.99528	-3.7 0.992922
HSapiens SPombe	-3 0.992318	-3 0.995851	-3.5 0.996678	-3.1 0.993828
MMusculus RNorvegicus	-2.5625 0.998962	-2.5 0.998516	-3.1 0.99382	-2.6 0.995405
MMusculus SCerevisiae	-3.5 0.988797	-3.5 0.979656	-4.05 0.988992	-3.45 0.991396
MMusculus SPombe	-3 0.993706	-3 0.996363	-3.5 0.995556	-3.2 0.991827
RNorvegicus SCerevisiae	-2.625 0.994812	-2.75 0.985235	-3.275 0.987427	-2.675 0.993865
RNorvegicus SPombe	-2.375 0.996311	-2.5 0.966572	-3.05 0.986327	-2.35 0.992244
SCerevisiae SPombe	-3.125 0.988385	-3 0.990529	-3.5 0.992774	-3 0.994183

Table 3: Final temperatures and pBads calculated using each method presented in logarithmic space

	Combination	Binary Search	Regression	Regression After Adjustment
AThaliana CElegans	-5.75 4.5917e-08	-5.75 2.70048e-07	-4.725 0.00806706	-5.625 4.29242e-06
AThaliana DMelanogaster	-7.0625 1.43104e-22	-6.5 2.45403e-08	-5.175 0.00621014	-6.275 3.34104e-06
AThaliana HSapiens	-6.75 2.28108e-14	-6.5 1.3041e-08	-5.25 0.010524	-6.35 3.74646e-07
AThaliana MMusculus	-6.125 7.69164e-08	-6.25 2.04443e-09	-5 0.00692956	-6.1 8.92199e-09
AThaliana RNorvegicus	-5.375 3.60316e-09	-6.25 1.88754e-48	-4.4 0.00308475	-5.1 8.31146e-06
AThaliana SCerevisiae	-7.3125 2.74494e-11	-7 4.58673e-06	-5.5 0.00240705	-6.8 4.37941e-06
AThaliana SPombe	-5.75 2.62836e-10	-5.5 1.25752e-05	-4.7 0.00211918	-5.5 1.53181e-06
CElegans DMelanogaster	-6.0625 5.14558e-14	-5.75 3.63015e-07	-4.8 0.00429799	-5.7 3.55858e-06
CElegans HSapiens	-5.9375 5.80703e-10	-6.25 1.49721e-09	-4.725 0.00687654	-5.625 8.31792e-06
CElegans MMusculus	-5.8125 1.56712e-07	-5.75 1.19434e-07	-4.8 0.00254389	-5.6 5.75805e-06
CElegans RNorvegicus	-5.25 1.8245e-08	-5.25 9.82673e-08	-4.5 0.00141717	-5.1 4.39485e-06
CElegans SCerevisiae	-6.25 3.71943e-19	-6.25 3.67555e-18	-4.8 0.00185127	-5.8 5.09154e-08
CElegans SPombe	-5.5625 7.19651e-07	-5.75 5.20985e-10	-4.7 0.0024433	-5.4 3.43364e-06
DMelanogaster HSapiens	-7.125 2.58125e-07	-7.25 4.96369e-10	-5.8 0.00607291	-6.9 2.77127e-06
DMelanogaster MMusculus	-6.125 3.70875e-07	-6.25 1.46565e-09	-5 0.00763119	-6 3.65177e-06
DMelanogaster RNorvegicus	-5.5 6.06613e-08	-5.5 3.37675e-11	-4.5 0.00161646	-5.2 9.8184e-06
DMelanogaster SCerevisiae	-7.5 5.52419e-08	-7.5 2.01522e-07	-5.8 0.00925233	-7.3 1.26334e-06
DMelanogaster SPombe	-6.4375 7.9975e-42	-6.25 8.73967e-06	-4.725 0.00221608	-5.525 1.08507e-06
HSapiens MMusculus	-6 4.59143e-07	-6.25 5.46857e-08	-5 0.0102182	-6.1 3.38876e-07
HSapiens RNorvegicus	-6.0625 3.68926e-27	-5.75 2.91272e-12	-4.4 0.00406681	-5.2 6.56055e-06
HSapiens SCerevisiae	-7.6875 4.94201e-10	-7.5 1.05296e-08	-6 0.0109763	-7.2 7.60018e-06
HSapiens SPombe	-6.5625 4.35543e-08	-6.25 1.31382e-28	-4.8 0.00127587	-5.6 4.31454e-06
MMusculus RNorvegicus	-5.3125 1.27876e-09	-5.5 1.01462e-14	-4.4 0.00150542	-4.9 5.52623e-06
MMusculus SCerevisiae	-6 6.30372e-07	-6.25 4.2138e-08	-4.875 0.00350308	-5.975 1.04122e-10
MMusculus SPombe	-5.625 2.04889e-07	-6.25 3.38173e-25	-4.7 0.00173563	-5.6 3.80634e-06
RNorvegicus SCerevisiae	-5.625 8.8892e-16	-6.5 5.52696e-49	-4.25 0.002081	-5.15 3.63889e-06
RNorvegicus SPombe	-5 2.75583e-07	-5.25 4.28773e-11	-4.25 0.000600686	-5.05 1.74682e-06
SCerevisiae SPombe	-5.875 1.65356e-12	-6.75 1.59283e-79	-4.6 0.00254432	-5.3 6.81862e-06

Table 4: Execution time of each algorithm

	Combination	Binary Search	Regression
Ancylostoma ceylanicum Pan troglodytes verus	63.439	35.121	265.485
Barnadesia spinosa Cichorium intybus	90.041	49.119	403.319
Boechera stricta Cynoglossus semilaevis	74.826	49.014	322.936
Bos indicus Mouse cytomegalovirus	91.239	55.087	411.162
Caenorhabditis brenneri Cynoglossus semilaevis	73.669	43.55	324.94
Citrus clementine Caenorhabditis remanei	109.825	62.383	482.651
Clytia hemisphaerica Cucumis sativus	31.096	19.116	125.084
Coccomyxa sp C 169 Zingiber officinale	41.772	24.592	174.492
Cryptomeria japonica Brugia malayi	83.527	41.378	350.476
Cynoglossus semilaevis Brugia malayi	83.153	46.975	367.317
Erysimum cheiri Picea sitchensis	96.039	56.113	416.627
Gnetum gnemon Pisum sativum	64.558	36.386	281.607
Gossypium herbacium Rehmannia glutinosa	39.129	24.271	162.488
Juglans hindsii x Juglans regia Aquilegia caerulea	96.919	54.416	438.415
Lupinus luteus Cucumis sativus	31.147	19.205	124.23
Nasturtium officinale Zingiber officinale	41.586	24.577	172.767
Nicotiana sylvestris Nuphar advena	60.038	36.781	256.527
Oryza barthii Prunus armeniaca	50.498	32.177	212.194
Oryza brachyantha Salmo salar	117.444	72.224	485.236
Ostrinia nubilalis Eucalyptus globulus	41.583	24.633	167.968
Papilio xuthus Cynoglossus semilaevis	71.992	43.23	315.461
Penaeus monodon Manihot esculenta	96.684	53.935	404.961
Petunia x hybrida Cucumis sativus	30.926	18.138	129.216
Poecilia reticulata Antirrhinum majus	48.73	28.422	205.499
Rehmannia glutinosa Quercus robur	77.676	45.474	338.02
Rhodnius prolixus Cryptomeria japonica	48.472	30.411	201.824
Salvia miltiorrhiza Salmo salar	111.553	66.09	505.237
Schedonorus arundinaceus Choristoneura fumiferana	38.059	22.227	155.842
Solanum demissum Cochliomyia hominivorax	70.47	38.046	295.355
Sorghum propinquum Centaurea maculosa	88.708	45.883	372.822
Strongyloides ratti Dendroctonus ponderosae	30.683	18.91	122.184
Trifolium pratense Manihot esculenta	97.813	53.168	440.478
Zingiber officinale Dissostichus mawsoni	58.014	31.944	244.73

Table 5: Amount of samples taken from a network pairs TR relationship

	Combination	Binary Search	Regression
Ancylostoma ceylanicum Pan troglodytes verus	29	14	117
Barnadesia spinosa Cichorium intybus	29	14	122
Boechera stricta Cynoglossus semilaevis	29	17	117
Bos indicus Mouse cytomegalovirus	29	15	123
Caenorhabditis brenneri Cynoglossus semilaevis	29	15	122
Citrus clementine Caenorhabditis remanei	29	14	119
Clytia hemisphaerica Cucumis sativus	29	16	115
Coccomyxa sp C 169 Zingiber officinale	29	15	117
Cryptomeria japonica Brugia malayi	29	13	116
Cynoglossus semilaevis Brugia malayi	29	14	122
Erysimum cheiri Picea sitchensis	29	15	117
Gnetum gnemon Pisum sativum	29	14	122
Gossypium herbacium Rehmannia glutinosa	29	16	117
Juglans hindsii x Juglans regia Aquilegia caerulea	29	14	119
Lupinus luteus Cucumis sativus	29	16	114
Nasturtium officinale Zingiber officinale	29	15	116
Nicotiana sylvestris Nuphar advena	29	15	116
Oryza barthii Prunus armeniaca	29	16	115
Oryza brachyantha Salmo salar	29	16	115
Ostrinia nubilalis Eucalyptus globulus	29	15	114
Papilio xuthus Cynoglossus semilaevis	29	15	120
Penaeus monodon Manihot esculenta	29	14	116
Petunia x hybrida Cucumis sativus	29	15	119
Poecilia reticulata Antirrhinum majus	29	15	117
Rehmannia glutinosa Quercus robur	29	15	119
Rhodnius prolixus Cryptomeria japonica	29	16	116
Salvia miltiorrhiza Salmo salar	29	15	121
Schedonorus arundinaceus Choristoneura fumiferana	29	15	116
Solanum demissum Cochliomyia hominivorax	29	14	119
Sorghum propinquum Centaurea maculosa	29	13	119
Strongyloides ratti Dendroctonus ponderosae	29	16	114
Trifolium pratense Manihot esculenta	29	14	123
Zingiber officinale Dissostichus mawsoni	29	14	118

follow it exactly. This is not surprising because SANA is a probabilistic algorithm and the number of computations vary between runs. As shown in Table 6, it is found that the main cause of this is SANA's inaccuracy when approximating how many iterations it can complete in the allotted time. This is a possible explanation for why SANA might not finish with the expected pBad or temperature.

Table 6: Average iterations actually executed in 60 seconds divided by the amount of iterations that were estimated to run in 60 seconds.

AThaliana CElegans	0.953625
AThaliana DMelanogaster	0.969012
AThaliana HSapiens	1.02585
AThaliana MMusculus	0.848217
AThaliana RNorvegicus	0.874255
AThaliana SCerevisiae	0.94193
AThaliana SPombe	0.933489
CElegans DMelanogaster	0.959395
CElegans HSapiens	1.01422
CElegans MMusculus	0.867277
CElegans RNorvegicus	0.861556
CElegans SCerevisiae	0.851479
CElegans SPombe	0.879733
DMelanogaster HSapiens	1.0377
DMelanogaster MMusculus	0.958658
DMelanogaster RNorvegicus	0.897178
DMelanogaster SCerevisiae	0.976897
DMelanogaster SPombe	0.93015
HSapiens MMusculus	1.01534
HSapiens RNorvegicus	1.01324
HSapiens SCerevisiae	1.03023
HSapiens SPombe	0.970253
MMusculus RNorvegicus	0.904331
MMusculus SCerevisiae	0.998324
MMusculus SPombe	0.872089
RNorvegicus SCerevisiae	0.965019
RNorvegicus SPombe	0.887735
SCerevisiae SPombe	0.858651

As mentioned earlier, it is important to analyze a network pair's TP relationship before the simulated annealing actually begins. The new method is able to create a summary of a TP relationship and the end points of the lines are comparable to that of a linear regression from 100 samples. However, the slope of the middle line is much lower. This is problematic because the algorithm will not be adept at detecting invalid network pairs.

Acknowledgements

Thank you to Dr. Wayne Hayes for directing the SANA project. Also, gratitude is extended to the California Alliance for Minority Participation at UC Irvine for facilitating the classes that assisted in the writing of this paper.

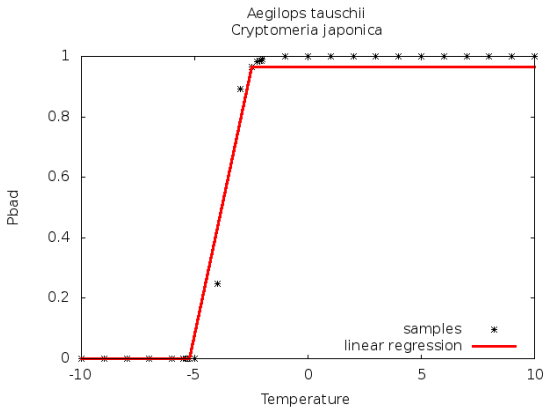

Figure 6: Linear Regression using 29 samples

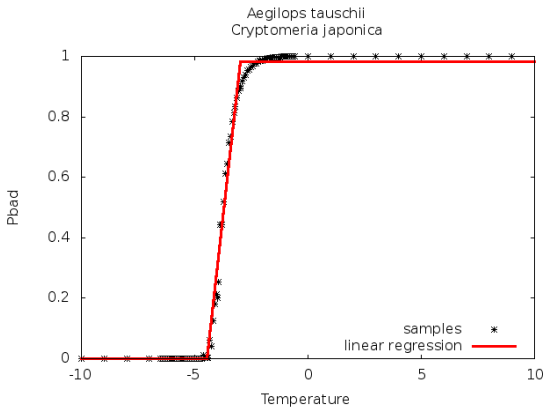

Figure 7: Linear regression using 100 samples

Citations

Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., ODonnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D. Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., and Tyers, M. (2013). The biogrid interaction database: 2013 update. Nucleic Acids Research, 41(D1), D816D823.

Hayes, W., Mamano, N. (2016). SANA: Simulated Annealing Network Alignment Applied to Biological Networks

Igor Jurisica, in a collaboration with Dr. Hayes, personal communication (2017). Toronto Network Collection