

**A Study of GeneWise with the *Drosophila Adh* Region**

Yi Mo, Moira Regelson, and Mike Sievers

Paracel Inc., Pasadena, CA

## ABSTRACT

GeneWise is one of the most accurate computer programs for gene finding, but unfortunately it is very computationally expensive. Paracel has accelerated GeneWise on its sequence analysis supercomputer, GeneMatcher™. In this study, the performance and scientific validity of Paracel GeneWise (PGW) were assessed by comparing PGW to software GeneWise (SGW), using the *Drosophila Adh* region as the benchmark. For an equivalent search, PGW running on GeneMatcher2 achieved a speed 2735 times faster than that of SGW running on a single 700-MHz Pentium III processor, yielding effectively the same results. A search was performed for all Pfam hits in the *Adh* sequence, comparing PGW to a heuristically accelerated GeneWise (HAGW) approach called "HalfWise." HalfWise uses BLASTX to select potential Pfam hidden Markov models (HMMs) for further analysis with the more computationally expensive GeneWise. The PGW approach had a sensitivity and specificity up to 87% and 88%, respectively, for identifying Pfam HMM hits, compared to 59% and 93% with the HAGW approach. The exceptional speed and proven scientific validity of Paracel GeneWise make it an indispensable tool for annotations in the genomic era.

## INTRODUCTION

With the onslaught of floods of genomic DNA sequence data, including the human genome (Venter et al., 2001; International Human Genome Sequencing Consortium, 2001), the need for computational tools to rapidly and accurately annotate genomes is ever more pressing. Among various software programs for gene finding and genome annotation in large DNA sequences, GeneWise (Birney and Durbin, 2000; <http://www.sanger.ac.uk/Software/Wise2>) stands out as one of the most accurate (Guigo et al., 2000). GeneWise is a protein-homology based program using hidden Markov models (HMMs) for finding genes in genomic DNA sequences. By incorporating a protein profile-HMM and a model of DNA splice sites, GeneWise finds the best gene structure prediction and, simultaneously, the alignment of the genomic sequence to the protein profile-HMM or protein sequence. However, GeneWise is a very computationally expensive dynamic program, so it is not yet used very widely for large-scale genome annotations. In order to make GeneWise a more practical tool in the genomic age, Paracel has implemented and vastly accelerated the algorithm on GeneMatcher™, a supercomputer for biological sequence analyses.

To assess the performance of Paracel GeneWise (PGW) relative to software GeneWise (SGW), and its scientific validity relative to heuristically accelerated GeneWise (HAGW), we have evaluated each approach with a genomic DNA sequence contig of about 2.9 Mb from the *Drosophila Adh* region. This region has been extensively studied and was used in the Genome Annotation Assessment Project (GASP) (Reese et al., 2000). In this evaluation, we have focused on finding all Pfam (Bateman et al., 2000) protein profile-HMMs that occur in the *Adh* genomic sequence, a study similar to one done by Birney and Durbin (2000). Pfam is a database of protein profile-HMMs and multiple sequence alignments for protein domains and families. It is widely used for genome annotation because of the functional information that can be inferred from similarities to protein domains and families.

## METHODS

### *Searches with HAGW*

For HAGW searches we used a Perl script, `halfwise.pl` (included in the Wise2 software package: <http://www.sanger.ac.uk/Software/Wise2>), to reduce the computational cost of running GeneWise. The HalfWise approach consists of two steps. In the initial step, HalfWise uses BLASTX to search DNA sequences against a protein database (`halfwise.db`) consisting of the seed alignments of models in the Pfam 5.5 database. The results of this search are used to select Pfam HMMs with possible hits. In step two, these models are used with the more computationally intensive GeneWise database search and sequence alignment algorithm. To further reduce the computational requirements of our test, the DNA sequence was split into segments of 100 kb. Except for the terminal piece, which contained the last 100 kb of the *Adh* sequence, these segments did not overlap.

We performed two HAGW searches using "`halfwise.pl`" with the database of *Adh* sequence segments. We performed the initial BLASTX filter step once and selected potential HMMs corresponding to the hits with E-values < 0.001. These HMMs were then used in two GeneWise searches. Both searches were run with the flag "`-kbyte`" to use all the available physical memory and the "`-pthread`" option to use multiple CPUs where applicable. The flags "`-pretty -gff`" were used to ease analysis and 200 alignments were requested.

For the first search, we used the "6LITE" algorithm for both the database search and sequence alignment steps of software GeneWise. 6LITE is the simplified version of the GeneWise algorithm implemented on GeneMatcher™ (Regelson et al., 2000). To ensure an equivalent comparison between SGW and PGW we used the flags "`-splice flat -alg 6LITE -aalg 6LITE`." This search allowed a direct comparison of speed and scores between SGW and PGW.

In the second search, a genome annotation approach, we used the 6LITE algorithm for the database search step but performed the sequence alignment step with the 623L algorithm (flag "`-aalg 623L`"). The 623L algorithm has a looping mode to obtain multiple high-scoring pairs (HSPs) for an HMM query on a single DNA sequence (Birney and Copley, 1999). The score threshold of the database search step was 20 bits (the default) for both HAGW searches. For the results of the second HAGW search we used a bit score threshold of 11 to select hits for further analysis, since the alignment algorithm occasionally produces lower bit scores than the search algorithm. The start and end positions of each hit on an *Adh* segment were corrected to their positions in the original, intact *Adh* sequence. For hits on the same DNA strand with overlapping corrected positions, only the hit with the highest bit score was kept for further analysis.

### *Searches with PGW*

For comparison to the results of the first HAGW search, we performed a PGW search of the *Adh* segment database with the HMMs selected from HAGW's BLASTX filter step. The search was run with a score threshold of 20 bits and the "`-no_max_introns`" option, which, like SGW, allows introns of unlimited size. The limitation of intron length is a feature introduced to PGW by Paracel to penalize hits with

introns longer than a specified size. This helps eliminate insignificant hits with long introns. This limit is only imposed on the hardware search, not at the post-processing stage. Therefore hits with introns longer than the limit, but with significant matching exons as well, can still be reported. The results of this PGW search and the first HAGW search are directly comparable.

To compare to the HAGW approach for genome annotation using the looping mode algorithm 623L, we ran a PGW search to find multiple HSPs (MHSP) in the *Adh* segment database. We used all the models in Pfam 5.5, a score threshold of 20 bits, and set the parameter "max\_intron\_length" to 1000 bases. Due to the parallel nature of the GeneMatcher™'s technology, it is not feasible to implement the looping mode of 623L directly. Thus, Paracel has implemented a combination of hardware and software techniques to obtain multiple HSPs. Setting the "max\_intron\_length" parameter to 1000 bases served two purposes: it minimized the possibility of joining two tandem repeated domains with a spurious intron and dramatically improved the speed performance of the PGW MHSP algorithm.

The first step in finding MHSPs with PGW is conducted inside the GeneMatcher™ hardware. During a search with a query HMM, the GeneMatcher™ returns multiple hits in a data sequence if the scores of the hits are local maxima and above the score threshold. To meet the criteria for a local maximum, the running maximum alignment score must have fallen below a lower threshold since the last reported hit. The hits reported by the GeneMatcher™ are then filtered to remove those with end points lying too close to the end points of other, higher scoring hits. The alignments of the filtered multiple HSPs are then computed according to a greedy algorithm. The highest scoring hit is aligned first and subsequent alignments may not overlap with previous ones. Hits with endpoints lying along the alignment of higher score hits will be discarded. To further filter random hits to HMMs with low-complexity regions, we restricted the hits of the PGW MHSP search to those that matched complete HMMs, and those with scores of at least 30 bits and at least 40 amino acids of an HMM matched.

## RESULTS AND DISCUSSIONS

### *Speed*

The BLASTX step of the HAGW searches took 3 hrs 58 mins on a single 700 MHz Pentium III processor of a Dell computer with 4 GB physical memory running the SunOS 5.7 operating system. This entailed searching 30 100kb DNA sequences from the split *Adh* sequence against the "halfwise.db" database of 116326 protein sequences. The BLASTX results yielded 96 protein profile-HMMs for use in the SGW search against the split *Adh* sequences, the second phase of the HAGW search. The SGW step using the 6LITE algorithm for both database search and sequence alignment took 68 hrs 22 mins running "genewisedb" in Wise2.1.23c on the Dell computer described above, and 16 hrs 50 mins on a Dell computer with 8 700 MHz Pentium III processors and 8 GB memory.

In contrast, it took only 1.5 mins for the corresponding PGW search and alignment on a 9-board GeneMatcher2 and a Dell post-processor with 8 700-MHz Pentium III CPUs and 4 GB physical memory. The PGW search, therefore, ran at a speed 2735 times faster than SGW running on a single CPU and 673 times faster than SGW on 8 CPUs.

With PGW, we were able to go a step further and search the *Adh* segment database with all the models in Pfam 5.5. The total time to search and align was just 9 minutes. For comparison, we can extrapolate from our results with the 96 selected HMMs to the time this search would take in SGW. The 96 selected HMMs have a total of 24573 nodes, whereas the 2478 HMMs in Pfam 5.5 have 594340 nodes in total. Since SGW's search time scales linearly with the number of HMM nodes, the search time for all of Pfam 5.5 would be approximately 24 times that with the selected 96 models. Analysis of SGW shows ~96% of the time spent on searching and ~4% on alignment, so the search alone for SGW on a single CPU would take more than 66 days. This indicates a speed-up factor of more than 10600 for PGW over SGW. By the same calculations, the extrapolated SGW search time on 8 CPUs would be more than 16 days, which would mean a speed-up of more than 2600 times (Table 1).

**Table 1.** Speed comparison between SGW and PGW<sup>1</sup>

	96 selected HMMs <sup>2</sup>	Whole Pfam 5.5 <sup>3</sup>
SGW 1-CPU (min.)	4102	95409 <sup>4</sup>
SGW 8-CPU (min.)	1010	23492 <sup>4</sup>
PGW (min.)	1.5	9
Speed-up over 1-CPU	2735	10601
Speed-up over 8-CPU	673	2610

<sup>1</sup>SGW was running “genewisedb” with flags “-init default -cut 20 -aln 200 -splice flat -alg 6LITE -aalg 6LITE -kbyte 3600000” on a single 700-MHz Pentium III CPU, and with the same flags except for “-kbyte 5700000 -pthread” on 8 700-MHz CPUs. PGW was running on a 9-board GeneMatche2 with a post-processor with 8 700-MHz Pentium III CPUs.

<sup>2</sup>These 96 HMMs were selected from the results of the BLASTX search of the split *Adh* sequences against “halfwise.db”. Only models represented by hits that have E-values less than 0.001 were selected. There are 24573 nodes in total for these models.

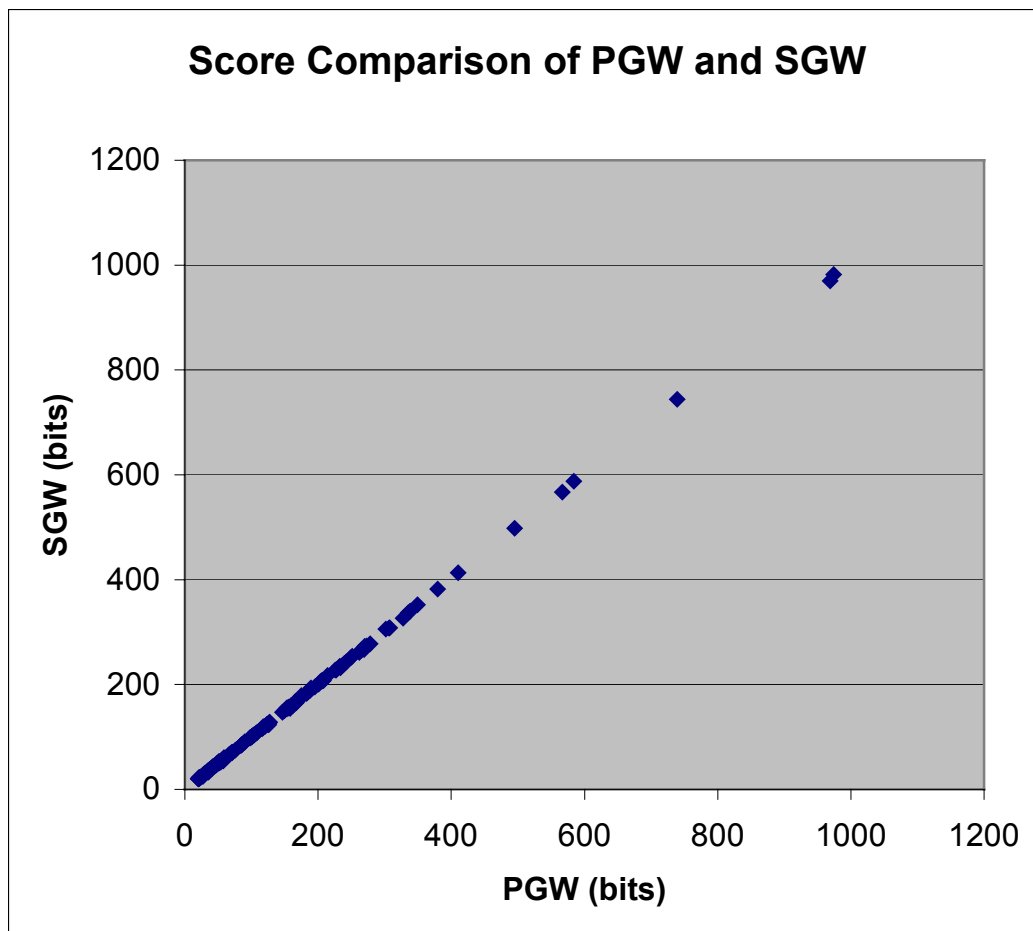
<sup>3</sup>There are 2478 HMMs with 594340 nodes in Pfam 5.5.

<sup>4</sup>This time was extrapolated from the SGW time for the 96 selected HMMs, which used ~96% of the time for searching and ~4% of the time for alignment, assuming that the alignment time would be the same and searching time ~24 times (594340 / 24573) longer.

It is interesting to note that PGW seems to be highly efficient when a search is scaled-up. For the search with all Pfam 5.5 HMMs, with about 24 times as many nodes as the smaller search, the total search time was only about 6 times as long. On the other hand, a general purpose computer with 8 times as many CPUs and twice as much physical memory only achieved a speed-up of about 4 times for SGW, indicating an efficiency of about 50%.

### ***Score Comparison***

The SGW search, using the 6LITE algorithm and the 96 HMMs selected from the BLASTX results, found 144 hits with 84 of the HMMs. The comparable PGW search found hits with the same 84 HMMs: 143 of the hits found by SGW and 6 additional hits. The bit scores of the 143 common hits were compared between SGW and PGW. The overall average difference in the scores was only 0.98 bits, or 0.67%, a result comparable to what Regelson et al. found from their study of a search with 3 HMMs and 988 hits (Fig. 1). After a closer look at the alignments of all the hits, the alignments of 134 hits were found to be exactly the same between SGW and PGW, while the alignments of the remaining 9 hits were somewhat different. The overall average score difference of the 134 identical hits was 0.97 bits, or 0.65%, essentially unchanged. The score and alignment differences were both probably due to scaling differences between SGW and PGW. It is likely that these slight score differences are also responsible for the 1 missing and 6 additional PGW hits compared to the SGW hits, since all those hits had scores very close to the threshold of 20 bits.



**Figure 1.** Score comparison of PGW and SGW. 143 hits of 84 HMMs found by both PGW and SGW are compared. The average difference of the scores is only 0.98 or 0.67%.

### *Assessment of Genome Annotation Approaches of HAGW and PGW*

The second HAGW search using the 6LITE algorithm for database search and the 623L algorithm for sequence alignment was considered as the HAGW approach for genome annotation. It took 67 hrs 26 mins to finish on a single CPU (15 hrs 51 mins on 8 CPUs), in addition to 3 hrs 58 mins for the BLASTX step. By contrast, on a 9-board GeneMatcher2 with an 8 700-MHz CPU postprocessor, the comparable PGW MHSP search with all of Pfam 5.5 took less than 14 mins to finish.

To analyze the search results of both HAGW and PGW, we checked the hit lists for multiple HMMs matching the same region of the *Adh* genomic sequence. For hits overlapping more than 9 bases, only the highest scoring one was kept. The final results of the HAGW and PGW searches comprised two lists of length 241 and 361, respectively, of HMM hits on the *Adh* genomic sequence. These lists were used to assess the scientific validity of both the HAGW and PGW approaches for genome annotation.

A standard dataset, std3, used by the GASP (<http://www.fruitfly.org/GASP>) was used here as the reference for the assessment. The std3 dataset was based on exhaustive and careful annotations of the *Adh* region (Ashburner et al., 1999; <http://www.fruitfly.org/about/pubs/ashburner99.html>), and it contains 222 annotated gene products. To confirm the accuracy of PGW and HAGW hits, we used both the alignment of the std3 protein sequences to the *Adh* region and the results of a HMMER search with all of Pfam 5.5 against the std3 protein sequences. The HMMER search was performed with a bit score cutoff of 0 and an E-value cutoff of 0.01, but hits with negative bit scores and E-values less than 0.001 were included if supported by a BLASTX hit.

### *Sensitivity and Specificity*

Of the 222 protein-encoding genes in the std3 dataset, the two GeneWise approaches considered found a combined total of 122 genes with at least one verified hit to a Pfam model. The PGW approach detected 120 of the genes in the std3 dataset, including 94 of the 96 genes found with the HAGW approach. In total, 55% of the genes in the *Adh* region were found to be similar to at least one Pfam model, with 89 Pfam models having at least one occurrence in the std3 dataset. This percentage of matching Pfam domains in the *Adh* region is 5 percentage points higher than that in human chromosome 22 (Dunham et al., 1999) probably due in part to the use of a later version of the Pfam database and to the sensitivity of the GeneWise algorithm.

The PGW approach found 361 hits with 103 distinct Pfam HMMs, while the HAGW approach identified 241 hits for 80 models. By inspecting the GeneWise alignments between the HMMs and the genomic sequence and by comparing the HMM alignments to the std3 protein dataset, we were able to confirm 292 of the PGW hits as true positives. For the HAGW approach, only 198 hits were confirmed as true positives. To determine false negatives, we considered verified hits missed by one approach and found by the other. In addition, we determined from the HMMER search of the std3 dataset that an additional 38 potential true hits were missed by both approaches.

The PGW approach found 115 confirmed hits missed by the HAGW approach, so

we estimate the total number of false negatives for HAGW to be 153. The PGW approach only missed 11 of the verified hits found by HAGW, so we estimate the number of false negative hits for PGW to be 49. Thus, the sensitivity and specificity for detecting Pfam hits with PGW were estimated to be 86% ( $292 / (292 + 49)$ ) and 81% ( $292/361$ ), respectively. Similarly, the sensitivity with the HAGW approach is 56% ( $198 / (198 + 153)$ ), and the specificity is 82% ( $198/241$ ) (Table 2).

**Table 2.** Sensitivities and specificities of the PGW and HAGW genome annotation approaches at the Pfam hit level<sup>1</sup>

	PGW	HAGW
Sensitivity	86%	56%
Specificity	81%	82%
Sensitivity corrected <sup>2</sup>	87%	59%
Specificity corrected <sup>2</sup>	88%	93%

<sup>1</sup>Sensitivity was calculated with  $(\text{true positives})/(\text{true positives} + \text{false negatives})$ , and specificity was calculated with  $(\text{true positives})/(\text{true positives} + \text{false positives})$ , i.e.  $(\text{true positives})/(\text{total hits identified})$ .

<sup>2</sup>Here retroviral Pfam HMM hits were considered as true positives in a broader sense of genome annotation that includes retrotransposons.

Interestingly, both the HAGW and PGW approaches identified 25 hits for three retroviral Pfam HMMs (rve, PF00665; rvp, PF00077; rvt, PF00078). These are probably true hits for retrotransposons, but not for genes. For genome annotations it may be desirable to include information about retrotransposons, which represent a major feature of many genomes. So the sensitivity at the Pfam hit level including retrotransposons for the PGW and HAGW approaches may be corrected to 87% ( $(292 + 25) / (292 + 25 + 49)$ ) and 59% ( $(198 + 25) / (198 + 25 + 153)$ ), respectively. And the specificity for the PGW and HAGW approaches may be corrected to 88% ( $(292 + 25) / 361$ ) and 93% ( $(198 + 25) / 241$ ), respectively (Table 2).

Apparently, the PGW approach has a much higher sensitivity and a slightly lower specificity than the HAGW approach. The PGW approach identified many hits for HMMs that were not selected from the BLASTX step of the HAGW approach. Among those HMMs a few, such as LRR (PF00560) and EGF (PF00008), are short motifs and have highly duplicated positive hits: 57 and 22, respectively. The HAGW approach failed to find those hits, so its sensitivity at the hit level was therefore much lower. On the other hand, the PGW approach also identified a few extra short, low-complexity HMMs and several false hits such as protamine\_P1 (PF00260, arginine rich) and metalthio (PF00131, cysteine rich). These hits somewhat reduced the specificity of the PGW approach. Overall, the PGW approach showed a 28-30% increase in sensitivity, at the expense of a 1-5% decrease in specificity, relative to the HAGW approach. Therefore, the PGW approach seems to be a more accurate way to directly annotate genomic DNA sequences with Pfam.

### *Effects of Algorithmic Differences*

When we compared the hits found by the PGW and HAGW approaches, we found 169 hits in common. These included 138 hits matching to the std3 proteins and 25 hits matching to retroviral protein domains. Repeated hits with 17 HMMs account for 86 of these 169 hits in common. 41 additional PGW hits correspond to 53 HAGW hits, but do not match exactly. Of the PGW hits, 12 were split differently by the HAGW approach and appear as 27 HAGW hits. There were 6 additional PGW hits that were split relative to 3 HAGW hits. These differences seem to be due to the use of "max\_intron\_length" of 1000 bases in PGW and to the algorithmic differences between the PGW and HAGW approaches for identifying multiple HSPs. The remaining 23 hits have alignment differences of varying degrees, which may result from the intron length limitation in the PGW approach and slight score differences from scaling.

The major algorithmic difference between the PGW and HAGW approaches is that the HAGW approach uses the looping mode algorithm 623L to find MHSPs as opposed to the heuristic greedy algorithm used by the PGW approach. It is of interest to look closer at those discrepant hits in order to assess how these algorithms compare to each other. The looping mode of HAGW is intended to identify multiple repeated hits for a HMM by maximizing the sum of the scores of the repeated hits, so it does not look for the single best alignment as the non-looping mode does. On the other hand, the PGW approach uses a greedy algorithm to select from a set of potential HMM hits.

For most repeated hits, both the HAGW and PGW approaches worked equally well, as indicated by the 86 hits found by both approaches. However, there were 12 PGW hits corresponding to 27 HAGW hits. These hits might represent the tendency of the looping mode used by the HAGW approach to generate fragmented hits at gapped regions in some alignments (Fig. 2). 5 additional hits had shorter alignments

**A**

Alignments of top-scoring domains:

1-4|300000: domain 1 of 1, from 5042 to 4484: score 107.1, LogLenNorm = 34

```

PF02137 Adenos      1 ALGTGNKCVSGPDEHISLNGTVLNDCHAEIVARRGLLRFLYSQLLLFNS 49
  +LG G+KC+ +   ++ +NG+ LND+HAE++ARRG+LRFLY +L +++
  SLGCGTKCIGE--SKLCPNGLIILNDSHAEVLARRGFLRFLYQEL-KQDR
1-4|300000          5042 tcgtgaatagg--tactcagcacagtcggggcgccgtcctctcgc-acga 4907
  ctgggcagtg--catgcagtttaacacattcgggttgtaaataa--aaag
  gtacaagctaa--cattctcccgttctcgggattattctacggc-ggta

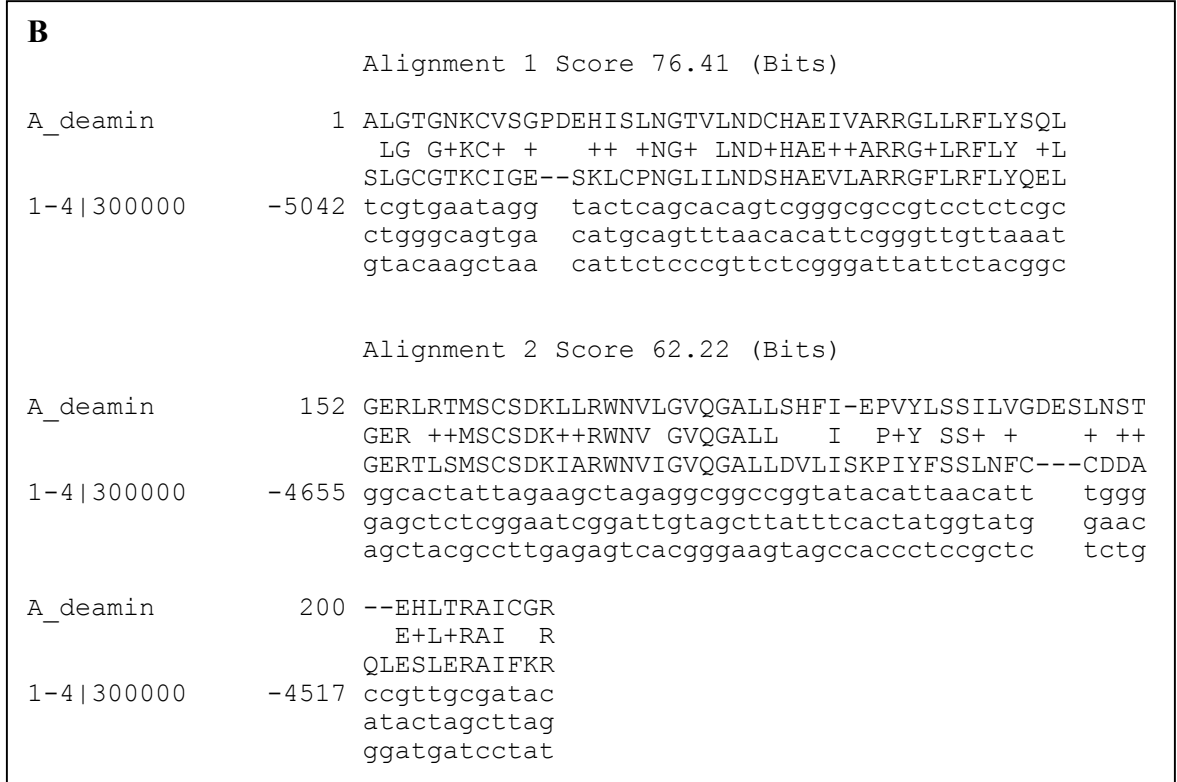
PF02137 Adenos      50 KNQATPEDSIFEKAKEGKLLRLKPNVSFHLIYINTAPCGDARIDSKLES 98
  IF + + + V FH + + PCGDA I +
  -----IFHWNSTLS-TYDMDEHVEFHFLSTQTPCGDACILEEEQP
1-4|300000          4907 -----atctaaaca-atgagggcgtctcaacactgggtacgggcc 4790
  -----ttagagctg-caataaataatattgcaccggacgttaaaac
  -----cttgccgat-gccgtgcccgcctgcaagttgtccctgagaa

PF02137 Adenos      99 DTSSNEETEDKHIPVFRKARQGLRRTKIENGEGETVPVAVESSDAVQTD 147
  ++ + + + + +G + + + +QT
  AARA-----KRQR-LDEDSE-----MVYTGAKLIS-DLSDDPMLQT-P
1-4|300000          4787 ggcg-----accc-cgggtg-----agtaggacaa-gcagggcatca-c 4685
  ccgc-----agag-taaaca-----ttacgcattg-atgaacttac-c
  gggc-----acgg-gtgtcg-----gctgtcaaat-ctcttagggg-a

PF02137 Adenos      148 GIL-----LGERLRTMSCSDKLLRWNVLGVQGALLSHFI-EPVYLSSIL 190
  G+L GER ++MSCSDK++RWNV GVQGALL I P+Y SS+
  GALRTKPRGERTLSMSCSDKIARWNVIGVQGALLDVLISKPIYFSSLN
1-4|300000          4682 ggccaacgcggcactattagaagctagagggcgccggtatacattaaca 4538
  gctgcacgggagctctcggaatcggattgtagcttatttactatggta
  atgagggcatagctacgccttgagagtcacgggaagtagccaccctccgc

PF02137 Adenos      191 VGDESLNST--EHLTRAICGR 209
  + + ++ E+L+RAI R
  FC---CDDAQLESLERAIFKR
1-4|300000          4535 tt---tgggccgttgcgatac 4484
  tg---gaacatactagcttag
  tc---tctgggatgatcctat

```



**Figure 2.** An example of the alignment of split SGW hits. A) The PGW alignment to a hit of the adenosine-deaminase (editase) domain (A\_deamin, PF02137) . B) The SGW alignments of the corresponding hits produced by the 623L algorithm.

in the HAGW results than their counterparts in the PGW results and appear to be fragmented (Table 3). Hence HAGW seems to produce about 13% ((27+5)/241) fragmented hits. On the other hand, there were only 6, or ~2%, PGW hits fragmented relative to two HAGW hits. We expect some amount of fragmentation in PGW because of the limitation of introns to 1000 bases. This limitation may force exons on either side of a long intron to be reported as distinct hits. Apparently, both the PGW and HAGW approaches can produce fragmented hits. The percentage of fragmented hits might be underestimated, since some of the 86 hits found by both PGW and HAGW could be fragmented as well.

**Table 3.** Numbers of split HAGW hits and split PGW hits

Split HAGW hits (2 or more fragments) <sup>1</sup>	Corresponding PGW hits
27	12
Shorter fragmented HAGW hits <sup>2</sup>	Corresponding PGW hits
5	5
Split PGW hits <sup>3</sup>	Corresponding HAGW hits
6	3

<sup>1</sup>Two or more HAGW hits correspond to a single PGW hit. The HAGW hits were fragmented by the looping mode of HAGW.

<sup>2</sup>One fragmented HAGW hit here corresponds to one PGW hit, which has a longer alignment.

<sup>3</sup>The PGW hits were split with PGW MHSP approach and the “max\_intron\_length=1000”.

The "max\_intron\_length" parameter is a feature introduced in PGW. We have examined the effect of this option in Table 4. Upon comparing hits that differed between the two approaches, we found that PGW had missed two immunoglobulin domains (Ig, PF00047) found by HAGW. This appears to have been due to the intron size cutoff, since the fragmented domains could be found by lowering the score threshold of the search. While the intron size limit may occasionally lead to false negative results, so can hits with excessively long introns. In the HAGW results, 10 true hits were excluded from the HAGW hit list because they overlapped with higher-scoring hits with incorrect large introns. HAGW found an additional 4 hits that were correct, but were reported with long and incorrect introns (> 10kb). These hits were eliminated from the hit list because the

large intron size caused overlap with other higher-scoring hits. In total, about 9% ((4 + 10)/153) of the HAGW false negative hits resulted from spuriously large intron size and 4% (2/49) of the PGW false negative hits resulted from the limitation of intron size. There was also one HAGW hit reported with an intron larger than 10kb, which the corresponding PGW hit did not have. The PGW hit without the intron seemed correct, as verified by the corresponding HMMER hit the std3 dataset. Thus, the introduction of the "max\_intron\_length" option to PGW appears to benefit its overall sensitivity and accuracy of genome annotation.

**Table 4.** Effects of unlimited intron size in HAGW<sup>1</sup> and the parameter "max\_intron\_length=1000" in PGW

HAGW hits with large introns overlapped with other higher scoring hits	laminin_EGF, laminin_B, MutS_N, 7tm_3	
True HAGW hits with large introns	FAD_Gly3P_dh	P450
Eliminated hits <sup>2</sup>	-	zf_C2H2 (6) WD40 (4)
False negative hits with PGW	Ig (2)	

<sup>1</sup>HAGW does not have the option to limit the maximal intron size, as opposed to the "max\_intron\_length" option implemented in PGW.

<sup>2</sup>The hits were eliminated by the true HAGW hit that has large introns. The numbers in parentheses represent how many multiple copies were eliminated.

### ***Potential Annotation Errors in the std3 Dataset***

In the course of this study, five possible annotation errors in the std3 dataset have emerged (Table 5). Both the PGW and HAGW approaches found a strong match to an ABC transporter domain (ABC\_tran, PF00005) (Fig. 3). It is next to a gene, *BG:DS00797.5*, containing an ABC\_tran domain as well, so it might be a part of a duplicated gene. Another two potential annotation errors consist of two Peptidase\_M2 domains (PF01401) found by both approaches close to a gene, *Ance*, also containing a Peptidase\_M2 domain. This also suggests that these two additional Peptidase\_M2 domains may be parts of duplicated genes. The last potential annotation error is a missing 7tm\_3 domain (PF00003), which partially matches a gene, *BG:DS00929.6*, but

with a much higher score from the PGW result than that from the HMM search of the std3 dataset (149.6 vs. -15.8 bits). This hit was found by HAGW but was not included in the hit list because of two large introns that overlapped with other higher-scoring hits. Interestingly, during GASP Birney and Durbin (2000) noted annotation oversights for 10 other exons. These potential errors in the std3 dataset beg further investigation and verification with biological experiments.

## CONCLUSION

In this study, the performance and scientific validity of Paracel GeneWise (PGW) were assessed by comparing PGW to software GeneWise (SGW) and to a heuristically accelerated GeneWise (HAGW), using the *Drosophila Adh* region as the benchmark. For an equivalent search of 96 Pfam HMMs against the chopped *Adh* sequence, PGW running on GeneMatcher2 achieved a speed 2735 times faster than SGW running on a single 700 MHz Pentium III processor. The results were essentially the same between PGW and SGW, with score difference of 0.65% on average and almost all alignments the same.

The scientific validity of PGW and HAGW was assessed with the identification of all Pfam 5.5 hits for the *Adh* region, using a well-annotated standard dataset as reference. The strategies for carrying out this task with PGW and HAGW differed in two aspects. First, the HAGW approach, "halfwise", used BLASTX to select

**Table 5.** Potential annotation errors in the std3 dataset<sup>1</sup>

Pfam Hit name	Position on <i>Adh</i> <sup>2</sup>		std3 counterpart
ABC_tran, PF00005	280181	280839	-
Peptidase_M2, PF01401	403474	405385	-
Peptidase_M2, PF01401	409497	412395	-
7tm_3, PF00003 <sup>4</sup>	1519294	1520373	<i>BG:DS00929.6</i>

<sup>1</sup>Only PGW hits are listed.

<sup>2</sup>The start and end positions of the hits on the 2.9-Mb *Adh* sequence. A hit is on the reverse strand if the number on the left is larger than the number on the right.

<sup>3</sup>Corresponding hits found by the HAGW approach are fragmented.

<sup>4</sup>Corresponding hit from HAGW had two large introns and overlapped with other higher-scoring hits, thus was eliminated.

potential Pfam models for the more computationally expensive GeneWise to use, while the PGW approach could afford to search all the Pfam HMMs. This was a significant

difference that caused the sensitivity of the PGW approach (86-87%) to be higher than that of the HAGW approach (56-59%) by 28-30%, with a cost in lowered specificity (PGW 81-88% vs. HAGW 82-93%) of only 1-5%. The second difference between the PGW and HAGW approaches was how multiple hits on a single database sequence were identified, with HAGW using the looping mode implementation and PGW using a heuristic greedy algorithm with a limit for the intron size. While for most multiply occurring hits these two strategies worked equally well, the difference seemed to cause a few discrepancies in the final results, with both PGW and HAGW fragmenting some domains.

Taken together, the study of GeneWise with the *Drosophila Adh* region has shown that PGW dramatically accelerates SGW, while it still maintains the same if not superior scientific validity as HAGW does. Therefore, PGW clearly presents itself as a powerful tool for genome annotations in the genomic age. Besides Pfam HMMs, GeneWise can also use proteins directly for gene predictions. Although the gene coverage of Pfam 5.5 is only about 55%, the gene coverage of a non-redundant full protein set should be much higher. Only PGW can carry out a search of that size in a reasonable time frame, giving the best results for genome annotations. As the sizes of both Pfam and protein databases such as SwissProt increase, PGW will truly distinguish itself as the indispensable tool for genome annotations.

#### **ACKNOWLEDGMENT**

We would like to thank Drs. Ewan Birney, Joe Borkowski, Cecilie Boysen and James Candlin for useful discussions.

```

Alignment 1 Score 151.91 (Bits)

ABC_tran      1 GEVLALVGPNGAGKSTLLKLISSLLPPTTEGTTILLDGARDLR
E+++L+G+NGAGK+T +++I G + +I++ G +++
KEITVLLGHNGAGKTTMMNMIMGRDS---NSIRFLG-HPMH
1-3|200000    80181 agaagccgcagggaaaaaaaaaagagt  atactcg ccac
aatctttgaagcgaccttatttggac  actgttg acta
ggttcggcctataaacggcgagtgca  catcccc tagc

ABC_tran      42
L:I[att]
SKLKERLERLRKNIGVVFQDPTL
+ +++ R IG+++Q+ ++
1-3|200000    80292 ATGTTGATT Intron 1 CAGTg E---RDVASYRHLIGFCPQHSVF
<2-----[80294:80349]-2> a gatccagattgtgcaagtt
g ctgcctccgtactggcctc

ABC_tran      66 FPNVELTVRENIAFGLRSLGLSKDEQRARLKKAGAEELLERLGLGYDH
+T + + F+++L G+ + ++R++ A+E L++LGL +
MS--YMTCHQHLEFFAQL-RGACRSDARDW-----ADEKLKKLGLS--D
1-3|200000    80411 aa taatccctgttgcc cggctcggcgt gggacaacgca g
tg atcgaatatattcat ggcggcacgag caaataatgtg a
gc cgatcgcggctcgg gatcactactg atgatgggatc t

ABC_tran      115 LLDRRPGTSLGGQKQRVAIARALLTKP
+++++++LSGG+K+R+++ A++ +
KRNEFGKNLSGGMKRRLSLGIAIAGNT K:K[aaa]
1-3|200000    80528 acagtgaattggaaacttcgagaggaaAAGTAAACT Intron 2
agaatgaatcggtaggtctgtctcgac <2-----[80611:80673]
gacaccgtgtacgggtaagtcctctcc

ABC_tran      142 LLLLEDEPTAGLDPASRAQLLELLRELRQQGGTVLLITHDLDDLDR
+++LDEP +GLD++SR++L+++L LR+ +++VL++TH+++++
IVILDEPSSGLDINSRRELWDILLNLRK-EKAVLVTTHYMEEAEV
1-3|200000    80671 CAGAagacggcttgtgaatccgctgaccacca gaggcgaactaggggg
-2> ttttaaccgtatacggatgatttatga aactttccaataacat
cactcaaagagctcgttgggccattacg ggctgcccgccggcgc

ABC_tran      188 LADRILVLEDG
L D I++L++G
LGD TICILANG
1-3|200000    80807 cgaatatgag
tgactgttcag

```

**Figure 3.** An example of a potential annotation error in the std3 dataset. ABC\_tran (PF00005) has a hit that was identically found by both PGW and HAGW but not present in any genes in the std3 dataset.

## REFERENCES

- Ashburner, M. et al. 1999. An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster* Adh region. *Genetics* **153**:179-219.
- Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. 2000. The Pfam protein families database. *Nucleic Acids Res.* **28**:263-266.
- Birney, E. and Copley, R. 1999. Wise2 documentation (version 2.1.20 stable). <http://www.sanger.ac.uk/Software/Wise2>.
- Birney, E. and Durbin, R. 2000. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**:547-548.

- Guigo, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. 2000. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.***10**:1631-1642.
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
- Reese, M.G., Hartzell, G. Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. 2000. Genome annotation assessment in *Drosophila melangaster*. *Genome Res.***10**:483-501.
- Regelson, M., Arehart, A., Gill, T., Borkowski, J.A., Slater, G., and Birney, E. 2000. Genewise port to the Paracel GeneMatcher™. The 12th International Genome Sequencing and Analysis Conference (GSAC 2000) Miami Beach, Florida.
- Venter, J.C. et al. 2001. The sequence of the human genome. *Science* 291: 1304-1351.