



Contents lists available at ScienceDirect

## Fungal Genetics and Biology

journal homepage: [www.elsevier.com/locate/yfgbi](http://www.elsevier.com/locate/yfgbi)Analysis and prediction of gene splice sites in four *Aspergillus* genomes

Kai Wang, David Wayne Ussery, Søren Brunak\*

Center for Biological Sequence Analysis, Department of Systems Biology, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark

## ARTICLE INFO

## Article history:

Received 8 May 2008

Accepted 25 September 2008

Available online xxxxx

## Keywords:

*Aspergillus*

Splice site predictor

Artificial neural networks (ANNs)

Bioinformatics

Web server

## ABSTRACT

Several *Aspergillus* fungal genomic sequences have been published, with many more in progress. Obviously, it is essential to have high-quality, consistently annotated sets of proteins from each of the genomes, in order to make meaningful comparisons. We have developed a dedicated, publicly available, splice site prediction program called NetAspGene, for the genus *Aspergillus*. Gene sequences from *Aspergillus fumigatus*, the most common mould pathogen, were used to build and test our model. Compared to many animals and plants, *Aspergillus* contains smaller introns; thus we have applied a larger window size on single local networks for training, to cover both donor and acceptor site information. We have applied NetAspGene to other *Aspergilli*, including *Aspergillus nidulans*, *Aspergillus oryzae*, and *Aspergillus niger*. Evaluation with independent data sets reveal that NetAspGene performs substantially better splice site prediction than other available tools. NetAspGene will be very helpful for the study in *Aspergillus* splice sites and especially in alternative splicing. A webpage for NetAspGene is publicly available at <http://www.cbs.dtu.dk/services/NetAspGene>.

© 2008 Elsevier Inc. All rights reserved.

## 1. Introduction

The filamentous fungal genus *Aspergillus* comprises over 182 species (Archer and Dyer, 2004). Several genomes have been published so far: *Aspergillus fumigatus*, a common human pathogen (Pel et al., 2007); *Aspergillus nidulans*, a convenient genetic model organism which have contributed much to classical genetics (Galagan et al., 2005); *Aspergillus oryzae*, an important organism for production of fermented foods mainly in northeast Asia (Machida et al., 2005), *Aspergillus niger* which is a key organism which has long been used for production of citric acid and other industrial enzymes (Pel et al., 2007), and currently another two species closely related to *A. fumigatus*: *Neosartorya fischeri* and *Aspergillus clavatus* (Fedorova et al., 2008). *A. fumigatus* has been widely studied since it is a potentially deadly human pathogen and a major allergen. It is commonly found growing on organic debris in the soil, disperses non-sexual spores called conidia in the atmosphere, relying on disturbances of the environment and strong air currents (Latge, 1999). Exposure to *A. fumigatus* can cause an allergic response in sensitive individuals. All humans breathe in at least several hundred *A. fumigatus* conidia per day. However, the conidia are eliminated efficiently by the innate immune system (Chazalet et al., 1998). *A. fumigatus* is particularly harmful to humans whose immune system has been compromised by disease, drug therapy or genetic conditions. More importantly, *A. fumigatus* is an opportunistic

pathogen of transplant patients, AIDS patients, and other immune compromised individuals. The incidence of invasive aspergillosis in immunocompromised hosts can be as high as 50%, and the mortality ranges from 50% to 100% (Denning, 1998). The cost to treat diagnosed cases of invasive aspergillosis was around \$633 million in the USA in 1996; \$64,500 for each case (Dasbach et al., 2000). Only two antifungal drugs were produced before 2001 (Denning et al., 2002).

*Aspergillus niger* and *A. nidulans* can also cause infections. *A. fumigatus* is the most common *Aspergillus* species causing invasive disease, responsible for approximately 90% of human infections. It was believed that the virulence of *A. fumigatus* is due to the immunosuppression or genetic deficiency of the host rather than unique fungal determinants (Tekaiia and Latge, 2005). Thermotolerance is a trait critical to thrive in mammalian bodies. *A. fumigatus* is able to grow over 50 °C (Denning, 1998), and it grows faster than any other airborne fungus at 40 °C (Tekaiia and Latge, 2005).

Little was known about the *A. fumigatus* genomic structure before the genome sequence of clinical isolate Af293 became available (another isolate A1163 were published recently (Fedorova et al., 2008)). The Af293 isolate genome consists of a 29.4 Mbp sequence, eight chromosomes, and 9926 genes (Nierman et al., 2005). In this genome, gene sequences are often split into coding regions (exons), and intervening sequences (introns). Cells must precisely excise the noncoding intron sequences, and ligate the coding exon sequences by spliceosomal processing. Splice sites are sequence signals at two ends of introns. The recognition of those sequences is an important initial step of splicing. Mutation of the splice site sequences inactivates often splicing (Sharp, 2005).

\* Corresponding author. Fax: +45 45 93 15 85.

E-mail addresses: [wangkai@cbs.dtu.dk](mailto:wangkai@cbs.dtu.dk) (K. Wang), [dave@cbs.dtu.dk](mailto:dave@cbs.dtu.dk) (D.W. Ussery), [brunak@cbs.dtu.dk](mailto:brunak@cbs.dtu.dk) (S. Brunak).

In the past twenty years, a number of gene finders and splice site predictors have been developed, but there are very few available options in particular for *Aspergillus*. Even for predictions on higher eukaryotic species, the performance is still far from good, especially when considering alternative splicing where multiple transcripts exist for one gene (Brent, 2008; Mathe et al., 2002). A general approach towards the identification of functional sequence elements is to combine the outputs from several different prediction tools. Compared to complete gene finders, which are helpful for identifying genomic gene locations and different kinds of gene features (start codon, stop codon, UTR, CDS and so on), splice site predictors focuses typically on the putative donor and acceptor sites where donor and acceptor sites are predicted separately. That is, the donor and acceptor sites do not have to be paired like in gene finders, so this provides an opportunity to list multiple donor sites paired with one acceptor site, or one donor site with multiple acceptor sites relevant in the case of alternative splicing. So a set of splice site candidates can thus be used to design probes for microarray analysis of gene expression or alternative splicing in a genome-wide manner. Furthermore, accurate splice site signals with higher confidence values can be quite useful in the precise prediction of genes.

We present a novel and species-specific predictor of splice sites trained on *A. fumigatus* sequences based on neural network algorithms. It performs very well on the *Aspergillus* species when compared to other tools. The prediction results on the NetAspGene web server are also shown graphically and in GFF3 format.

## 2. Materials and methods

Artificial neural networks (ANNs) have been used successfully for splice site prediction e.g. in genes in humans and plants (Brunak et al., 1991; Hebsgaard et al., 1996). By adjusting the weights in the networks, ANNs are able to generalize beyond the training data. It is obviously essential to test the predictive performance on independent data sets.

### 2.1. Data preparation

The quality of the underlying data set is extremely critical. Unfortunately, there are often many errors appearing in biological databases and in genome annotation. We tried to filter out predicted *Aspergillus* genes, and then used very strict criteria to check and validate the data quality and reliability before training the algorithms.

The *A. fumigatus* genome (isolate Af293) was downloaded from NCBI. The genomic sequence consists of 9,923 genes with 18,282 pairs of splice sites (99.7% of them are GT–AG pairs). Due to the presence of either orthologs from different species or paralogs from gene families, many genes have in practice more than one copy. To assess the unbiased predictive ability of an artificial neural network, the data used for training and the data for testing must be non-similar. Therefore, homology reduction based on BLAST alignments at the amino acid level was performed. After creating a similarity list from the BLAST output, the Hobohm algorithm was used to set up a cutoff for the highest allowed similarity. The Hobohm algorithm is a robust selection algorithm to extract as many as possible high quality non-homologous examples from BLAST alignments (Hobohm et al., 1992). We did not consider the alignments from BLAST with lengths less than 10 nt; and also not lengths greater than 80 nt with BLAST percent identities smaller than 24.8%; as well as lengths between 10 and 80 nt when the BLAST percent identities were smaller than  $290.15 \cdot \exp(-0.562 \cdot \log(x))$ , where  $x$  is the length. After removing redundancy in this way, only about half the genes (4867) remained.

Furthermore, we only kept the well-annotated genes, and discarded all hypothetical and predicted genes. Last, the data set was not allowed to have any logical errors. Several criteria were used to validate the correctness of the remaining genes. The optimal gene was mainly defined by: intron number exceeding one; open reading frame size should be a multiple of 3 nt; and should contain correct start codon and stop codon. Finally, 1243 non-redundant genes with well-known and clear annotations were identified as the data set to use for neural network training. From this highly non-redundant set, we randomly selected 840 genes to train, 270 genes to test, and did leave 133 genes out for final evaluation. Compared to many other organisms, only limited amounts of EST and cDNA information is available for *Aspergillus*, and this fact possibly bias the set of 1,243 genes we used after filtering, and we still cannot be certain that this set fully represents the diversity of *Aspergillus* gene structure. This issue can only be solved when more and more genes are confirmed by experiments.

Comparing with humans and the plant *Arabidopsis*, *Aspergillus* has very different exon-intron structure, with an average exon size 400 nt and an average intron size 100 nt. In contrast, the average size for human exons is 150 and 740 nt for introns, while *Arabidopsis* has averages of 179 and 146 nt, respectively. This suggests that the architectures of the neural networks possibly should be different, and that they should be trained to find a novel combination different from other organisms.

### 2.2. Neural networks

The networks used in this study were of the multi-layer error back propagation type with three layers: an input layer, a hidden layer, and an output layer. The input was a segment of nucleotide sequences with an uneven window size in terms of nt. The output from the networks came from just one unit giving a value between 0.0 and 1.0, which was used to represent a category assignment for the middle nucleotide in the input window to assign if this nucleotide was predicted as coding or noncoding or if this nucleotide was predicted as a splice site or non-splice site. The entire gene structure was used as input for the neural networks to train. So, the exon and intron length distributions, as well as all other splice site patterns, have been included into our network training. In fact in a neural network with a large window the weights can adapt to situations where different length properties couple to different nucleotide sequence properties. This means that the prediction power goes far beyond schemes which just take a length distribution into account when scoring.

Many neural networks for coding/non-coding, donor sites, and acceptor sites were combined into one ensemble predictor (Brunak et al., 1991; Hebsgaard et al., 1996). Firstly, the separated donor and accept networks with different window sizes and hidden units were examined and combined: windows from 3 to 61 nt and hidden units from 0 to 100. Networks with 37 nt as the window size gave the best performance for donor sites; networks with 41 nt window size perform well for acceptor sites. To further enhance the predictions, we picked up 10 differently initialized architectures with different training window sizes and hidden units, and calculated the average as the prediction output individually for donor or acceptor sites. Next step, in order to obtain good performance for both small and large exons and introns, we did not just use only one optimal window size for the coding prediction networks, instead we averaged over networks with different window sizes between 101–401 nt after testing a large number of combinations for different networks. The extremely large window sizes for coding prediction cover both donor and acceptor sites of many exons, and gave better performance than single smaller window networks. Then, the predicted coding potential was used as a cut-off value for the prediction of local splice sites. The coding net-

works in NetAspGene are used to model the transitions from coding sequence to non-coding sequence in the vicinity of the predicted donor and acceptor sites according to a specific window size. That is, regions with abruptly increasing output activity of coding networks should enhance acceptor site assignment and suppress donor site assignment, while regions with abruptly decreasing coding activity should enhance donor site assignment. This was done by summing up the output activities to the left and right of the potential splice site, and calculating average difference on the two sides. If the output of a local network ensemble was greater than the average, the result is a positive splice site assignment. Finally, some post-prediction rules were also used to filter results in order to discard wrong and questionable splice sites. If a site is predicted in a uniformly low coding region, where the average difference of the coding predictions around this site can be close to zero in the oscillating regions. We will discard the false splice in the middle of uniformly predicted regions with strong assignment.

The Matthews correlation coefficient (CC) was used to quantify the performance and stop the training of the neural networks (Baldi and Brunak, 2001)

$$CC = \frac{T_p T_n - F_p F_n}{\sqrt{(T_n + F_n)(T_n + F_p)(T_p + F_n)(T_p + F_p)}} \quad (1)$$

where  $T_p$  (True positives) and  $T_n$  (True negatives) are the correctly predicted positives and negatives, and  $F_p$  (False positives) and  $F_n$  (False negatives) are the incorrectly predicted positives and negatives. Sensitivity and specificity were also used as means of quantifying and comparing the neural network performance with other prediction methods. The sensitivity ( $S_n = T_p / (T_p + F_n)$ ) of the prediction is defined as the probability of correctly predicting a positive site; the definition of specificity ( $S_p = T_p / (T_p + F_p)$ ) is the probability that a positive prediction is correct which is also known as the positive predictive value (PPV). A perfect prediction would give a correlation coefficient of 1, whereas a completely wrong one would give  $-1$ . The correlation coefficient takes both sensitivity and specificity into account. For example, a predictor that predicts every site to be positive, which is a predictor with 100% sensitivity but low specificity, would have a correlation coefficient of zero.

In order to compare with other *Aspergillus* species on NetAspGene, we downloaded the genomic sequences of *A. nidulans* EGSC A4, *A. oryzae* RIB40, and *A. niger* CBS 513.88 from NCBI (<http://www.ncbi.nlm.nih.gov/>), and extracted genes based on the gene structure annotations of the three genomes.

### 3. Results and discussion

#### 3.1. NetAspGene performance and comparison on the four *Aspergillus* species

The nucleotide composition around splice sites was analyzed qualitatively and visualized using sequence logos (Schneider and Stephens, 1990). The single nucleotide logo plots were generated for all splice sites of *A. fumigatus* (Fig. 1). Dinucleotides GT at the position 1 and 2 are the consensus sequence at donor sites; AG at the position  $-2$  and  $-1$  are the consensus sequence at acceptor sites. The other three *Aspergillus* species have strikingly similar signals around their splice sites. It is of interest to note that the polypyrimidine tract commonly seen in humans and plants, located before intronic acceptor sites, is not found for *Aspergillus*.

As described in the Methods section, we trained on 840 *A. fumigatus* genes, and tested on 270 genes using a combined network ensemble. The maximal correlation coefficient for donor sites was 0.80. When detecting 95% of the true donor sites, the combined approach makes only 0.18% false assignments. The maximal

correlation coefficient for acceptor sites was 0.76. When detecting 95% of the true sites, the combined approach makes 0.31% false assignments. In Fig. 2, the red curve is the prediction performance from the local network ensemble (small window sizes), while the green curve is the prediction performance for the combined network ensemble with different parameters after the post-prediction filtering.

NetAspGene was trained and modelled by *A. fumigatus* genes. In order to find out whether it can be generally used for other *Aspergillus* species, we have compared the genomic features and protein similarity of four *Aspergillus* species (Table 1 and Fig. 3). *A. oryzae* has the largest genome, but only 77% of the genes contain introns. *A. oryzae* genes contain the biggest average exon and intron sizes (460 and 120 nt). *A. fumigatus* has only 1.8 introns per gene, while *A. nidulans* have 2.7 introns per gene. *A. niger* has the smallest exon size. The four *Aspergillus* species have been claimed to depart several hundred million years ago, and their genomes therefore differ considerably. There are almost 3000 proteins that are closely related, or homologous between the genomes (Goffeau, 2005). *A. fumigatus* and *A. oryzae* share 70% identity, and each has 66–67% identity with *A. nidulans*, and this is comparable to the similarity between mammals and fish which diverged around 450 million years ago (Galagan et al., 2005).

The BLAST matrix shown in Fig. 3 illustrates the protein comparison between the four *Aspergillus* genomes. BLAST hits should be at least 80% of the gene lengths with an  $E$  value of  $1e-10$ . The diagonal shows the fraction of proteins that has homologous hits within the proteome itself; the fractions were shown in colors (grey or red). The comparison between genomes is shown in the colors grey or green. In general, these species clearly share many genes, and also to some extent codon usage. The high similarity of the four species indicates that NetAspGene can be generally used for other *Aspergillus* species. We randomly picked up 150 well-annotated and non-redundant genes from each of the other three *Aspergillus* species to test by NetAspGene. For *A. oryzae*, specificity of donor site prediction with high confidence was 78%; acceptor site 76%. For *A. nidulans*, specificity of donor site was 80%; acceptor site 83%. For *A. niger*, specificity of donor site was 79%; acceptor sites 81% (Table 2).

#### 3.2. Comparison to other predictors

In order to assess the performance of NetAspGene, we compared it with FSPLICE and GeneID, two species-specific splice site predictors for *Aspergillus*. We used 133 genes with 346 pairs of splice sites to evaluate the predictors, which had never been used for training or testing when developing NetAspGene. Based on the same sensitivity ratio of cutoff to the confidence values, we found the specificity of NetAspGene to be better both on donor and acceptor sites. GeneID, which uses Markov models to score sequences in a hierarchical scheme, has a specific version for *A. nidulans* (Parra et al., 2000). GeneID has in principle the best sensitivity but with 10 times more predicted donor sites than NetAspGene (14,909 vs. 1407 donor sites), and most of them are likely to be false positives (Table 3). On the other hand, FSPLICE predicts splice sites based on weight matrices model on many organisms including *Aspergillus* ([www.softberry.com](http://www.softberry.com)). FSPLICE predicts less potential candidates with a much lower specificity compared with NetAspGene, especially for acceptor sites (49% vs. 87%) (Table 3). We also tested the gene finder called FGENESH which applies hidden Markov model to predict gene exon-intron structures in many organisms (Salamov and Solovyev, 2000). It was believed to be the most accurate program for rice genes (Yu et al., 2002), and it has been widely used by researchers currently working on *Aspergillus* (Unsold and Li, 2006). It predicts with 85% specificity on donor sites; and 78% for acceptor sites, but it only covers 88% of the true

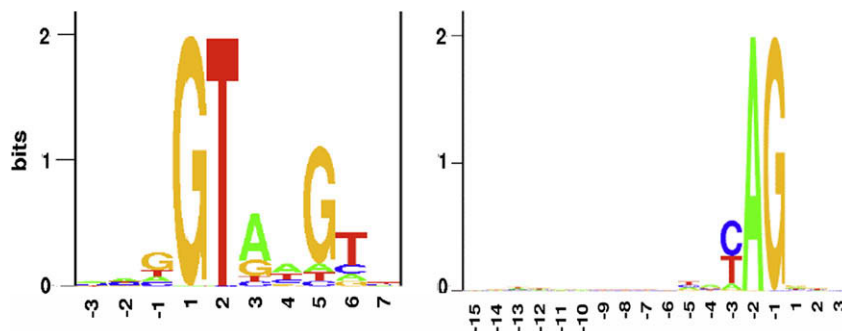


Fig. 1. The single nucleotide logo plots were generated for splice sites of *A. fumigatus*. The left plot is for donor sites; the right one is for acceptor sites.

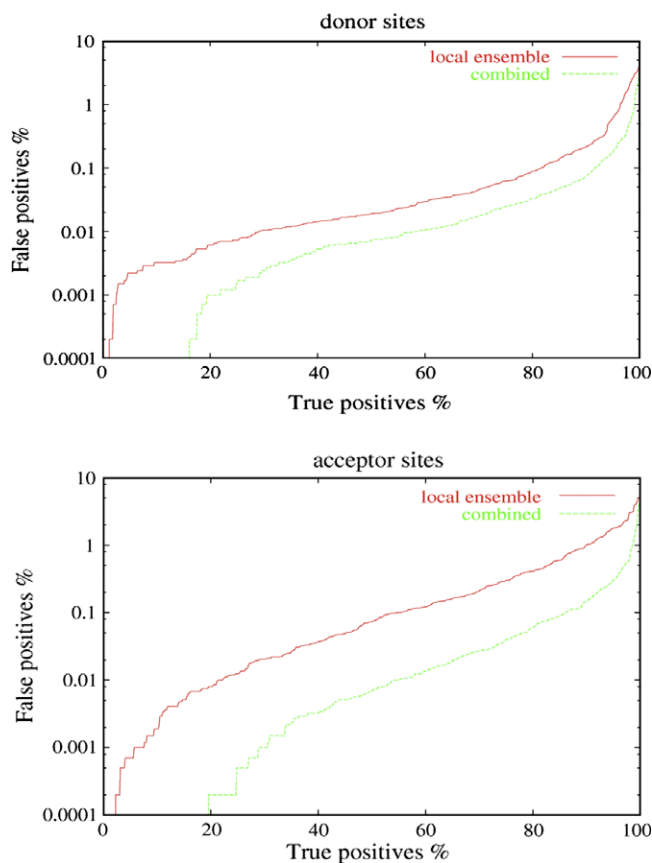


Fig. 2. Plots showing the network performance based on the single networks (in red) and combined networks (in green). The top plot is for donor site prediction; the bottom one is for acceptor site prediction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

Table 1  
Properties of the four *Aspergillus* genomes.

	<i>A. fum</i>	<i>A. nid</i>	<i>A. nig</i>	<i>A. ory</i>
Genome size (Mb)	28	30	34	37
Number of protein genes	9923	9541	13595	12074
Genes containing introns (%)	78	88	87	77
Average number of introns per gene	1.8	2.7	2.6	1.9
Number of splice site pairs	18282	25584	35046	23197
Average gene length (nt)	1732	1869	1567	1579
Average exon size (nt)	445	434	368	460
Average intron size (nt)	100	102	97	120

donor sites; and 80% of the true acceptor sites. It is not easy to decide which predictor eventually will be the best, since obviously

### Proteome Homology

Gene overlap > 80%, E-value < 1e-10

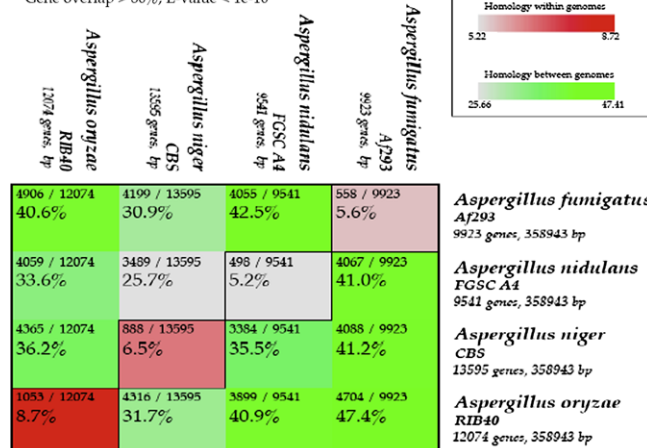


Fig. 3. Proteome homology between four *Aspergillus* species based on BLAST alignments made at the amino acid level.

Table 2

Performance of NetAspGene on *A. nidulans*, *A. niger*, and *A. oryzae* based on 150 randomly picked up genes from each species. 'D' represents donor site predictions, and 'A' represents acceptor site prediction.

	<i>A. nid</i>	<i>A. nig</i>	<i>A. ory</i>
(%) Specificity H (D)	80	79	78
(%) Specificity H (A)	83	81	76

Table 3

Performance comparison of NetAspGene with other *Aspergillus*-specific splice site predictors. All predictors were used on the direct strand and only considering GT as potential donor sites. In the top two rows, 'H' means predictions with high confidence values. In all rows, 'D' represents donor site predictions, and 'A' represents acceptor site prediction. The percentage values of FSPLICE and GeneID were the maximal percentages obtained at the same sensitivity ratio with NetAspGene.

	NetAspGene	FSPLICE	GeneID
(%) Specificity H (D)	80	66	65
(%) Specificity H (A)	87	49	74
(%) Sensitivity (D)	93	90	100
(%) Sensitivity (A)	90	89	91
(#) Predicted (D)	1407	1263	14909
(#) Predicted (A)	2370	3261	6162

there is no completely correct annotation available for evaluation. Many potential splice sites could be identified as true, but based on the well-annotated genes so far, it is safe to conclude that NetAspGene is the better option for splice site prediction on *Aspergillus*.

#### 4. Conclusion and availability

We have developed the currently best splice site predictor for *Aspergillus*. The method is made available through the WWW at <http://www.cbs.dtu.dk/services/NetAspGene>. The predictor can be combined with a gene finder to develop better annotation for *Aspergillus* genes. Furthermore, NetAspGene will be very helpful for the researchers who are interested only in splice sites and in particular alternative splicing. Predicting candidate splice sites by NetAspGene can be used to design probes for custom microarrays, for example with the aim to study alternative splicing.

#### Acknowledgments

We thank Niels Tolstrup and Hans Henrik Stærfeldt for technical support on the neural network algorithm and web server; we thank Peter Fischer Hallin, Thomas Nordahl Petersen, and Kristoffer Rapacki for additional assistance. This work was supported by the Danish National Research Foundation and the Natural Science Research Council.

#### References

- Archer, D.B., Dyer, P.S., 2004. From genomics to post-genomics in *Aspergillus*. *Curr. Opin. Microbiol.* 7, 499–504.
- Baldi, P., Brunak, S., 2001. *Bioinformatics—the machine learning approach*. MIT Press, Cambridge, Massachusetts.
- Brent, M.R., 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat. Rev. Genet.* 9, 62–73.
- Brunak, S. et al., 1991. Prediction of human messenger-rna donor and acceptor sites from the DNA-sequence. *J. Mol. Biol.* 220, 49–65.
- Chazalet, V. et al., 1998. Molecular typing of environmental and patient isolates of *Aspergillus fumigatus* from various hospital settings. *J. Clin. Microbiol.* 36, 1494–1500.
- Dasbach, E.J. et al., 2000. Burden of Aspergillosis-related hospitalizations in the United States. *Clin. Infect. Dis.* 31, 1524–1528.
- Denning, D.W., 1998. Invasive Aspergillosis. *Clin. Infect. Dis.* 26, 781–803.
- Denning, D.W. et al., 2002. Sequencing the *Aspergillus fumigatus* genome. *Lancet Infect. Dis.* 2, 251–253.
- Fedorova, N.D. et al., 2008. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. *PLoS Genet.* 4, e1000046.
- Galagan, J.E. et al., 2005. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature* 438, 1105–1115.
- Goffeau, A., 2005. Genomics—multiple moulds. *Nature* 438, 1092–1093.
- Hebsgaard, S.M. et al., 1996. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* 24, 3439–3452.
- Hobohm, U. et al., 1992. Selection of representative protein data sets. *Protein Sci.* 1, 409–417.
- Latge, J.P., 1999. *Aspergillus fumigatus* and Aspergillosis. *Clin. Microbiol. Rev.* 12, 310.
- Machida, M. et al., 2005. Genome sequencing and analysis of *Aspergillus oryzae*. *Nature* 438, 1157–1161.
- Mathe, C. et al., 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30, 4103–4117.
- Nierman, W.C. et al., 2005. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature* 438, 1151–1156.
- Parra, G. et al., 2000. GeneID in *Drosophila*. *Genome Res.* 10, 511–515.
- Pel, H.J. et al., 2007. Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nat. Biotechnol.* 25, 221–231.
- Salamov, A.A., Solovyev, V.V., 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* 10, 516–522.
- Schneider, T.D., Stephens, R.M., 1990. Sequence Logos—a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100.
- Sharp, P.A., 2005. The discovery of split genes and RNA splicing. *Trends Biochem. Sci.* 30, 279–281.
- Tekaia, F., Latge, J.P., 2005. *Aspergillus fumigatus*: saprophyte or pathogen? *Curr. Opin. Microbiol.* 8, 385–392.
- Unsold, I.A., Li, S.M., 2006. Reverse prenyltransferase in the biosynthesis of fumigaclavine C in *Aspergillus fumigatus*: gene expression, purification, and characterization of fumigaclavine C synthase FGAPT1. *Chembiochem* 7, 158–164.
- Yu, J. et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp indica). *Science* 296, 79–92.