First Pass Annotation of Promoters on Human Chromosome 22

Matthias Scherf,[1,3,4] Andreas Klingenhoff,[1,3] Kornelie Frech,[3] Kerstin Quandt,[3] Ralf Schneider,[1] Korbinian Grote,[1] Matthias Frisch,[3] Valérie Gailus-Durner,[1] Alexander Seidel,[1] Ruth Brack-Werner,[2] and Thomas Werner[1,3]

GSF-National Research Center for Environment and Health, [1]Institute of Mammalian Genetics; [2]Institute of Molecular Virology, Neuherberg, Germany; [3]Genomatix Software GmbH, Munich, Germany

The publication of the first almost complete sequence of a human chromosome (chromosome 22) is a major milestone in human genomics. Together with the sequence, an excellent annotation of genes was published which certainly will serve as an information resource for numerous future projects. We noted that the annotation did not cover regulatory regions; in particular, no promoter annotation has been provided. Here we present an analysis of the complete published chromosome 22 sequence for promoters. A recent breakthrough in specific in silico prediction of promoter regions enabled us to attempt large-scale prediction of promoter regions on chromosome 22. Scanning of sequence databases revealed only 20 experimentally verified promoters, of which 10 were correctly predicted by our approach. Nearly 40% of our 465 predicted promoter regions are supported by the currently available gene annotation. Promoter finding also provides a biologically meaningful method for "chromosomal scaffolding", by which long genomic sequences can be divided into segments starting with a gene. As one example, the combination of promoter region prediction with exon/intron structure predictions greatly enhances the specificity of de novo gene finding. The present study demonstrates that it is possible to identify promoters in silico on the chromosomal level with sufficient reliability for experimental planning and indicates that a wealth of information about regulatory regions can be extracted from current large-scale (megabase) sequencing projects. Results are available on-line at http://genomatix.gsf.de/chr22/.

The human genome sequencing project completed the first major milestone with the publication of most of the euchromatic part of human chromosome 22 (Dunham et al. 1999). The consortium identified a total of 545 genes using a careful approach, relying primarily on the mapping of experimental data such as cDNAs and EST clusters. In silico predictions were used to identify genomic data such as CpG islands and repetitive sequence contents.

The promoter of a gene is generally located in its 5′ region and contains vital information about gene expression and regulatory networks, including gene targets of individual transcriptional cascades/signaling pathways. However, cDNAs and EST clusters are often 5′ incomplete and thus do not provide reliable information about promoters. This and the scarcity of experimental data regarding promoters are probably the major reasons why no corresponding annotation for promoters was attempted.

It has not been possible thus far to predict polymerase II promoters in silico with sufficient specificity in the context of large genomic sequences. This problem was highlighted by the publication of the GASP project (Reese et al. 2000). We recently developed a new method called `PromoterInspector` (Scherf et al. 2000) to locate genomic regions of about 0.2 kb to 2 kb which contain or overlap with polymerase II promoters. We showed that `PromoterInspector` is capable of predicting promoter regions in sequences over 1 Mb in length with high accuracy. Approximately one-half of all predictions were confirmed by gene annotation and 43% of known promoters were detected (Scherf et al. 2000). These results indicate that `PromoterInspector` is able to overcome the problem of numerous spurious predictions in long sequences, which hampers all known promoter prediction tools [usually exceeding an error rate of 80% false positive matches (Fickett and Hatzigeorgiou 1997; Scherf et al. 2000]. Therefore, `PromoterInspector` appears well suited to the analysis of large genomic contigs such as those comprising the sequence of human chromosome 22.

The aim of the present study was to provide high-quality annotation of potential promoters on chromosome 22 specific enough to be useful for subsequent experimental design. `PromoterInspector` appeared to be an appropriate tool in achieving this goal, because the genomic regions located by `PromoterInspector` can be expected to contain complete promoters, which are generally less than 1 kb in length, and also because of the high reliability of `PromoterIn-`

[4]Corresponding author.
E-MAIL scherf@gsf.de; FAX 49 89–5490 8399.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.154601.

spector predictions (~50% correct predictions). Using chromosome 22 as an example, we demonstrate here that the annotation of genomic sequences can now be extended to include identification of promoter regions by in silico methods.

To our knowledge, the specificity of PromoterInspector has not been attained in promoter analyses of large genomic sequences to date, and no attempts to predict promoter regions of a whole human chromosome have been reported.

RESULTS

Experimentally Verified Promoters on Human Chromosome 22

To identify experimentally verified promoters on chromosome 22, we performed extensive searches in MEDLINE and GenBank. Because promoters are often referred to by a variety of expressions, we carried out both sequence-based BLAST searches (NCBI) (Altschul et al. 1990) and keyword-oriented text searches. We used genomic fragments containing 2 kb upstream and 500 bp downstream of the annotated gene starts as query sequences for BLAST. GenBank annotation as well as ENTREZ and MEDLINE were searched for entries containing the gene names and/or chromosome 22 annotation in order to include as many promoters as possible. Lastly, we mapped all entries of the Eukaryotic Promoter Database (EPD) (Perier et al. 2000) to the sequence of chromosome 22.

This approach yielded only 20 experimentally verified promoters of known genes on chromosome 22 (Table 1). We compared the location of the 20 promoters with the gene starts annotated by Dunham et al. (1999). In 18 cases, the experimentally verified promoters agreed very well with gene starts. However, two promoters (PLA2G6 and GGT1) were found to be located at a significant distance upstream of the annotated gene starts (12 kb, PLA2G6 and 20 kb, GGT1). We were able to map the 93 bp noncoding first exon of the PLA2G6 mRNA to the genomic sequence of chromosome 22 (ExonMapper, GEMS Launcher package). This exon was not included in the original annotation by Dunham et al. (1999), and the real gene start is located 12 kb upstream of the annotated gene start due to a large first intron. In the case of the GGT1 promoter, no continuous mRNA was available but the promoter sequence matches the chromosome 22 sequence over a stretch of more than 450 bp with just a single mismatch, showing no gaps at all. In summary, experimentally verified data are available for only a very few of the promoters on chromosome 22, even considering that we might have missed a few promoters due to unusual annotation. Therefore, large-scale promoter annotation requires appropriate in silico methods.

Table 1. Experimentally Verified Promoters on Human Chromosome 22

Accession no.	Gene	Strand	Annotated gene start relative to promoter	PromoterInspector predicted	CpG island predicted
L43122	*CONT*	+	within promoter	● (224 bp)	
X52828	*BCR*	+	within promoter	● (1088 bp)	● (1914 bp)
X84664	*MMP11*	+	18 bp downstream	● (216 bp)	● (591 bp)
AJ007494	*GGT1*	+	20,000 bp downstream[a]		
X72990	*EWSR1*	+	within promoter	● (1092 bp)	● (1284 bp)
M63420	*LIF*	−	12 bp upstream	● (236 bp)	● (473 bp)
AF129855	*OSM*	−	within promoter		
AF047576	*TCN2*	+	81 bp downstream		
AB016655	*LIMK2*	+	within promoter	● (452 bp)	● (520 bp)
S79779	*TIMP3*	+	120 bp downstream	● (624 bp)	● (886 bp)
S58267	*HMOX1*	+	within promoter		
EP11091[b]	*MB*	−	within promoter		
X63578	*PVALB*	−	within promoter		
X53093	*IL2RB*	−	9 bp upstream		
M87841	*H1F0*	+	within promoter		● (1018 bp)
AF115252	*PLA2G6*	−	12,000 bp downstream (93 bp exon 1 missing)	● (224 bp)	
EP11139[b]	*PDGFB*	−	931 bp downstream	● (860 bp)	● (2928 bp)
AF106656	*ADSL*	+	49 bp downstream		● (1059 bp)
D86746	*SREBF2*	+	within promoter	● (820 bp)	● (1476 bp)
M77378	*ACR*	+	within promoter		
Total:	20 genes			10 (50%)	10 (50%)

[a]Promoter sequences published separately, no continuous mRNA sequence available. All sequences are taken from GenBank/EMBL/EPD. See accession numbers for reference.
[b]EPD accession numbers.

Sequence Analysis and Promoter Region Predictions

PromoterInspector is an in silico method which is trained to predict the genomic context of polymerase II promoters. Details of the algorithm are as described earlier (Scherf et al. 2000). The application of PromoterInspector to chromosome 22 yielded 465 regions (minimum length 192 bp, maximum 2432 bp, average 555 bp, Table 2).

PromoterInspector predicts CpG island- as well as nonCpG island-associated promoter regions. Because 60% of human genes have distinctive CpG islands at their 5′ ends (Cross and Bird 1995) and the chromosome 22 sequence was found to be G + C rich (Dunham et al. 1999), we examined CpG island-associated predictions in more detail. Dunham et al. (1999) reported 553 CpG islands of which 543 were documented on the Web server of the Sanger Centre (http://www.sanger.ac.uk/cgi-bin/cwa/22cwa.pl). The minimum CpG island length is 400 bp, the maximum length is 10,000 bp and the average length is 1074 bp (Table 2).

Comparison of Promoter Region Predictions with Existing Annotation

PromoterInspector, as well as CpG islands predictions, yielded reasonable numbers of matches on chromosome 22 (PromoterInspector: 465 matches, CpG islands: 543 matches). The next step was to compare the results with existing gene annotations in order to determine whether the predictions were reliable.

The quality of the predicted regions was assessed on basis of the 5′ ends of the genes annotated by Dunham et al. (1999). We carried out a correlation analysis of all predicted promoter regions with annotated gene starts with the program package GenomeInspector (Quandt et al. 1996). The correlation analysis was done with respect to the different quality of gene annotation. We considered three groups of genes which were introduced by Dunham et al. (1999): (1) known genes (genes which are identical to human genes or protein sequences), (2) related genes (genes homologous, or containing a region of similarity, to gene or protein sequences from human or other species), and (3) predicted genes (sequences homologous to ESTs).

Table 2. Length of Predicted Promoter Regions and CpG Islands on Human Chromosome 22

	CpG	Promoter-Inspector
No. of predicted regions	543	465
Minimum region length	400 bp	193 bp
Average region length	1,074 bp	555 bp
Maximum region length	10,000 bp	2433 bp
Sequence coverage	583,645 bp	257,877 bp

Promoter regions were correlated with "known genes" and "related genes" within a region of 2 kb upstream and 0.5 kb downstream of the annotated gene starts (Fig. 1). In the case of the "predicted genes," the correlation peak was extended up to 6 kb upstream (Fig. 1B). The correlated promoter regions were considered "annotation-supported" promoter regions.

Table 3 summarizes the results of PromoterInspector predictions and CpG islands. The portion of regions correlated with gene is approximately the same for PromoterInspector predictions (38.7%) and CpG islands (39.4%). The numbers of the annotation-supported promoter regions and CpG islands might still be on the cautious side, especially in the case of EST-based gene annotation, where missing 5′ sequences can easily exceed 10 kb. In order to calculate the percentage of annotation-supported predictions, we set the total number of predictions obtained with each method to 100%. Dunham et al. (1999) identified a group of 134 "pseudo genes;" that is, sequences homologous to a known gene or protein sequence but with a disrupted open reading frame. Given a threshold of 2 kb, only six promoter regions predicted by PromoterInspector were correlated with a gene start in this group.

Because PromoterInspector predicts CpG island- as well as nonCpG island-associated promoters, we tried to improve the CpG island predictions by a filter approach: We considered only those CpG islands which overlapped with a PromoterInspector prediction. The filter approach resulted in 358 CpG islands, and 47.5% of them are correlated with an annotated gene start, as summarized in Table 4. In light of these results, the question arose as to whether this improvement could also be reached by filtering CpG islands with other in silico promoter prediction methods. We applied Promoter 2.0 (Knudsen 1999) and NNPP 2.1 (M. Reese, in prep.) and considered only those CpG islands where a promoter was predicted. Table 5 summarizes the results. NNPP 2.1 reduced the number of CpG islands without an improvement of predictions. Promoter 2.0 predicted a promoter in only 164 CpG islands, of which 52.2% are correlated with a gene start.

Finally, we considered PromoterInspector predictions which are not correlated with CpG islands. As can be seen in Table 1, 20% of the PromoterInspector regions which are correlated with an experimentally verified promoter are nonCpG island predictions. The correlation of nonCpG island PromoterInspector predictions with gene annotations is summarized in Table 6.

Gene Prediction Combined with Promoter Prediction

Dunham et al. (1999) applied GenScan (Burge and Karlin 1997), a program for identification of exon/

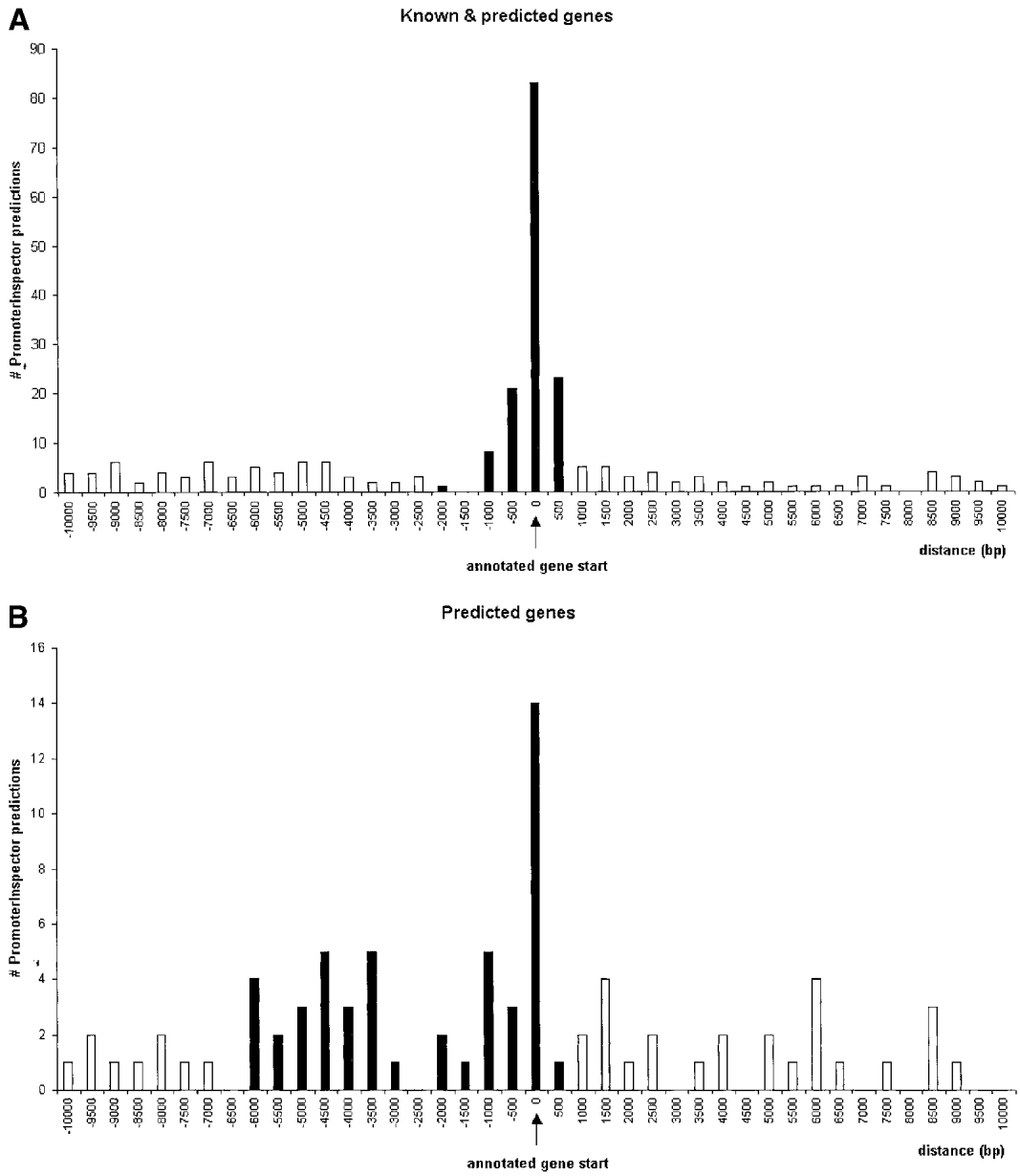

Figure 1 Correlation analysis of `PromoterInspector` promoter regions with annotated gene starts on chromosome 22 (+ strand shown). The *y*-axis indicates the total number of matches found in relative distance to the annotated gene start. Values on the *x*-axis with a negative sign mark distances to promoter regions which are located upstream of an annotated gene start, while positive values mark distances to promoter regions which are located downstream from an annotated gene start. The column at distance value 0 marks the number of promoter regions which overlap with an annotated gene start. The range accepted as tolerance is highlighted in black. (*A*), known and related genes as defined by Dunham et al. (1999). (*B*), predicted genes as defined by Dunham et al. (1999).

intron structures, to predict genes ab initio. A total of 817 `GenScan` predictions were obtained. Although 94% of the annotated genes were at least partially detected by `GenScan`, all exons were predicted correctly for only 20% of annotated genes. Because of these results, Dunham et al. (1999) stated that "... ab initio gene prediction cannot be used directly to annotate genes in human sequences."

We examined whether a combination of `GenScan` and `PromoterInspector` might improve the ab initio gene prediction. As a first step, we determined the set of composite predictions (i.e., `GenScan` gene predictions with a 5′ end within or at most 100 bp downstream from a `PromoterInspector` promoter region). A total of 92 `GenScan`/`PromoterInspector` predictions fulfilled this requirement. Again we used the gene annotations of Dunham et al. (1999) to estimate the reliability of these predictions.

Table 3. Correlation of Predicted Promoter Regions and CpG Islands with Gene Annotation on Human Chromosome 22

		CpG island 543 (100%)			PromoterInspector 465 (100%)		
Gene group	#	abs.[a]	rel.	O	abs.[a]	rel.	O
All genes	545	214	39.4%	5.8	180	38.7%	6.3
Known genes	247	126	23.2%	9.3	111	23.9%	11.1
Related genes	150	39	7.2%	4.9	28	6.0%	4.6
Predicted genes	148	55	10.1%	3.4	47	10.1%	3.6
Pseudo genes	134	9	1.7%	1.4	6	1.3%	1.2
Additional predictions		320	58.9%		279	60.0%	

(#) Number of annotated genes.
(abs) Absolute number of predicted promoter regions that were found near annotated genes.
(rel) Percentage of predicted promoter regions that were found near annotated genes.
(O) X-fold overrepresentation, i.e., number of predicted promoter predictions that were found near annotated genes divided by the number that would be expected to be located near these genes if the same number of posiitons were distributed randomly across the chromosome.
[a]Due to a small number of double scoring regions (a region is scored twice if it is supported by genes from different gene groups), the total number of annotation supported regions is not equal to the sum over the number of gene group-related annotation-supported promoter regions.

We found 11 composite predictions where the GenScan prediction did not overlap with an annotated gene, nor was the respective promoter region annotation-supported. Of the remaining 81 composite predictions, 49 (60.4%) had an annotation-supported promoter region and the respective GenScan predicted gene overlapped with the annotated gene. In 32 cases, the promoter regions were not annotation-supported but the respective GenScan prediction overlapped partially with an annotated gene.

Table 4. Correlation of CpG Islands Filtered by Predicted Promoter Regions with Gene Annotation on Human Chromosome 22

	CpG Islands correlated with PromoterInspector regions 358 (100%)			
Gene group	#	abs.[a]	rel.	O
All genes	545	170	47.5%	6.8
Known genes	247	105	29.5%	11.5
Related genes	150	29	8.1%	5.4
Predicted genes	148	41	11.4%	3.7
Pseudo genes	134	2	0.5%	0.4
Additional predictions		186	51.8%	

(#) Number of annotated genes.
(abs) Absolute number of predicted promoter regions that were found near annotated genes.
(rel) Percentage of predicted promoter regions that were found near annotated genes.
(O) X-fold overrepresentation, i.e., number of predicted promoter predictions that were found near annotated genes divided by the number that would be expected to be located near these genes if the same number of positions were distributed randomly across the chromosome.
[a]Due to a small number of double scoring regions (a region is scored twice if it is supported by genes from different gene groups), the total number of annotation supported regions is not equal to the sum over the number of gene group-related annotation-supported promoter regions.

From these results we concluded that composite predictions have a high chance (>50%) to correlate with true promoters. In addition, promoter regions appear to be useful markers for delineating the 5′ boundary of subsequences to be analyzed by GenScan. We could verify this for an example, the *SLCRA1* gene (Heisterkamp et al. 1995), which was originally not correctly predicted by GenScan. Using the annotation-supported promoter region predicted by PromoterInspector as a 5′ boundary, GenScan correctly recognized all exons of the *SLCRA1* gene.

In summary, our results suggest that the 11 additional composite predictions are more likely candidates for real genes than are isolated GenScan predictions, because GenScan and PromoterInspector independently identify different sequence features. Composite gene predictions might thus be useful as an in silico extension of the chromosome 22 annotation.

DISCUSSION

Promoters contain vital information about gene expression and regulatory networks, including gene targets of individual cascades/signaling pathways. To date, <5% of the promoters in chromosome 22 are known from experimental analysis.

We have shown that in silico promoter annotation of large-scale chromosomal sequences is feasible with a quality that is suitable for experimental design. Every second to third prediction of PromoterInspector and every second prediction of the PromoterInspector-filtered CpG island predictions can be shown to be correct. Although the annotation derived by our predictions is not complete (about every third annotated

Table 5. Correlation of CpG Islands Filtered by Promoter Prediction tools NNPP 2.1 and Promoter 2.0 with Gene Annotation on Human Chromosome 22

		NNPP–CpG 462 (100%)			Promoter 2.0–CpG 164 (100%)		
Gene group	#	abs.[a]	rel.	O	abs.[a]	rel.	O
All genes	545	177	38.3%	6.8	84	51.2%	7.1
Known genes	247	120	25.9%	10.3	55	33.5%	12.4
Related genes	150	33	7.1%	4.8	11	6.7%	4.2
Predicted genes	148	49	10.6%	3.6	21	12.8%	4.2
Pseudo genes	134	9	1.9%	1.6	5	3.0%	2.5
Additional predictions		276	59.7%		75	45.7%	

(#) Number of annotated genes.
(abs) Absolute number of predicted promoter regions that were found near annotated genes.
(rel) Percentage of predicted promoter regions that were found near annotated genes.
(O) X-fold overrepresentation, i.e., number of predicted promoter predictions that were found near annotated genes divided by the number that would be expected to be located near these genes if the same number of positions were distributed randomly across the chromosome.
[a]Due to a small number of double scoring regions (a region is scored twice if it is supported by genes from different gene groups), the total number of annotation supported regions is not equal to the sum over the number of gene group-related annotation-supported promoter regions.

gene was correlated with a predicted promoter region), it is, to our knowledge, the first successful large-scale prediction of promoter regions.

As can be seen in Table 3, PromoterInspector and CpG islands led to comparable numbers in gene start correlations. However, a significant difference between these two approaches is that the length of the PromoterInspector regions is, on average, one-half the size of the minimum length of CpG islands. Therefore, PromoterInspector predictions pinpoint gene starts with much more precision. In addition, the results in Tables 1, 4, and 6 show that PromoterInspector predicts CpG island- as well as nonCpG island-associated promoters. Our results show a bias of PromoterInspector predictions towards CpG islands. One reason for this is that ~60% of the promoter sequences which were used to train the PromoterInspector contained CpG islands. Since the training procedure focuses on the most common patterns in the training set (Scherf et al. 2000), it is clear that the prediction is biased towards C + G-rich patterns. However, our results show that the PromoterInspector approach is (to our knowledge) the only one able to predict promoter regions on the genome level with such a small sequence coverage (i.e., precision). To underline this statement, we compared PromoterInspector with the promoter prediction tools NNPP 2.1 (M. Reese, in prep.), TSSG (Solovyev and Salamov 1997), TSSW (Solovyev and Salamov 1997) and Promoter 2.0 (Knudsen 1999). These approaches focus on the detection of promoter elements like TATA and CAAT boxes rather than promoter regions. Since it is not possible to analyze whole chromosomes with these tools, we randomly extracted and analyzed 10 nonoverlapping sequences with a length of 50,000 bp from chromosome 22. From the obtained results we would expect 11,890 (TSSW), 14,963 (TSSW), 50,233 (Promoter 2.0) and 87,641 (NNPP 2.1) promoter predictions on the chromosome 22 sequence. Assuming that all promoters of the 545 annotated genes of chromosome 22 are included in these predictions, then

Table 6. Correlation of Predicted nonCpG-island Promoter Regions with Gene Annotation on Human Chromosome 22

	PromoterInspector regions not correlated with CpG Islands 85 (100%)			
Gene group	#	abs.[a]	rel.	O
All genes	545	10	11.8%	2.0
Known genes	247	5	5.9%	2.6
Related genes	150	1	1.1%	1.0
Predicted genes	148	4	4.7%	2.2
Pseudo genes	134	4	4.7%	0.8
Additional predictions		71	83.5%	

(#) Number of annotated genes.
(abs) Absolute number of predicted promoter regions that were found near annotated genes.
(rel) Percentage of predicted promoter regions that were found near annotated genes.
(O) X-fold overrepresentation, i.e., number of predicted promoter predictions that were found near annotated genes divided by the number that would be expected to be located near these genes if the same number of positions were distributed randomly across the chromosome.
[a]Due to a small number of double scoring regions (a region is scored twice if it is supported by genes from different gene groups), the total number of annotation-supported regions is not equal to the sum over the number of gene group-related annotation-supported promoter regions.

only every 20th to 140th prediction is expected to be correct. This is certainly not very useful for subsequent experimental design.

Annotated gene starts are not always useful for the identification of promoters because gene annotation might be 5′ incomplete, as suggested by Dunham et al. (1999). We found two examples in which experimentally mapped promoters were located more than 10 kb upstream of annotated gene starts, demonstrating that our methods were able to identify promoter-containing regions with high reliability. The approach of predicting promoter regions independent of gene annotations also provides a new way toward mapping of short first exons that are most frequently missed by both cDNA mapping and gene prediction (Dunham et al. 1999). A prediction upstream of the known gene sequence is very likely to represent the correct promoter and should also be a useful addition for annotated genes.

Reliable promoter prediction can also be used in a more general way to provide a biologically meaningful "chromosomal scaffold" for a variety of further analyses. For example, gene prediction tools such as GenScan perform much better when they are used on segments containing only one gene or at least starting with a gene. Therefore, the combination of promoter region prediction with gene prediction tools like GenScan is a promising way to enhance the specificity of de novo gene prediction. Our data already show a dramatic improvement in the amount of verified gene predictions obtained by simply combining the results of independent gene predictions with promoter predictions.

The next milestone in large-scale promoter analysis will be an in-depth in silico analysis of functional structures of promoters. Promoter function is defined by the specific arrangement of transcription factor binding sites. Promoters often contain subregions called transcriptional modules that are responsible for a specific transcriptional response of a promoter or a promoter group (Kel et al. 1999; Klingenhoff et al. 1999; Werner 1999). We previously showed that specific promoter modeling can yield functional insights into promoter organization in several cases (Frech et al. 1996, 1997, 1998), using a library of currently more than 100 computer models of transcriptional modules (GEMS Launcher). The module analysis will serve as the information base for ongoing research.

METHODS

Promoter Region Prediction

Promoter regions were predicted by PromoterInspector (Scherf et al. 2000). PromoterInspector predicts the genomic context of eukaryotic polymerase II promoter regions based on equivalence classes of IUPAC words. PromoterInspector is available on-line at http://genomatix.gsf.de/cgi-bin/PromoterInspector/PromoterInspector.pl.

Correlation Analysis

Correlation analysis was realized with the GenomeInspector software package (Quandt et al. 1996). GenomeInspector detects distance correlations between sequence elements on megabases of nucleotide sequences. The method is available on-line at http://www.gsf.de/biodv/software.html.

Exon Mapping

Exon mapping was achieved with the ExonMapper tool, which is an integral part of the GEMS Launcher software package (Genomatix Software GmbH; http://genomatix.gsf.de).

Promoter Mapping

We used the program FASTA (Pearson and Lipman 1988) with default parameters and the multiple alignment program DiAlign (Genomatix Software GmbH; http://genomatix.gsf.de/cgi-bin/dialign/dialign.pl) (Morgenstern et al. 1996) to find experimentally verified promoter sequences on the sequence of chromosome 22.

Gene Prediction

Gene prediction was carried out with the GenScan program (Burge and Karlin 1997) with default parameters. The method is available on-lineat http://CCR-081.mit.edu/GENSCAN.html.

Promoter Prediction

Promoter prediction was carried out with NNPP 2.1 (M. Reese, in prep.) TSSG (Solovyev and Salamov 1997), TSSW (Solovyev and Salamov 1997) and Promoter 2.0 (Knudsen 1999). All methods are available on-line as follows: NNPP 2.1, http://www.fruitfly.org/seq_tools/promoter.html; TSSG and TSSW, http://genomic.sanger.ac.uk/gf/gf.shtml; Promoter 2.0, http://www.cbs.dtu.dk/services/promoter.

ACKNOWLEDGMENTS

We thank Rudi Balling for critical reading of the manuscript. The help of Christian Mirschberger, Ina Stein and Elida Schneltzer is gratefully acknowledged. This work was partially supported by the BMFT Verbundprojekt FANGREB 514–4003–0311641 and DFG WE 2370/1–1.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F, Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Cross, C.H. and Bird, A.P. 1995. CpG islands and genes. *Curr. Opin. Genet. Dev.* **5:** 309–314.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human chromosome 22. *Nature* **402:** 489–495.

Fickett, J.W. and Hatzigeorgiou, A.C. 1997. Eukaryotic promoter recognition. *Genome Res.* **7:** 861–878.

Frech, K., Brack-Werner, R., and Werner, T. 1996. Common modular

structure of lentivirus LTRs. *Virology* **224:** 256–267.
Frech, K., Danescu-Mayer, J., and Werner, T. 1997. A novel method to develop highly specific models for regulatory DNA regions. *J. Mol. Biol.* **270:** 674–687.
Frech, K., Quandt, K., and Werner, T. 1998. Muscle actin genes: A first step towards computational classification of tissue specific promoters. *In Silico Biol.* **1:** 29–38.
Heisterkamp, N., Mulder, M.P., Langeveld, A., ten Hoeve, J., Wang, Z., Roe, B.A., and Groffen, J. 1995. Localization of the human mitochondrial citrate transporter protein gene to chromosome 22Q11 in the DiGeorge syndrome critical region. *Genomics* **29:** 451–457.
Kel, A., Kel-Margoulis, O., Babenko, V., and Wingender, E. 1999. Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J. Mol. Biol.* **288:** 353–376.
Klingenhoff, A., Frech, K., Quandt, K., and Werner, T. 1999. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15:** 180–186.
Knudsen, S. 1999. Promoter 2.0: For the recognition of Pol II promoter sequences. *Bioinformatics* **15:** 356–361.
Larsson, P.K., Kennedy, B.P., and Claesson, H.E. 1999. The human calcium-independent phospholipase A2 gene multiple enzymes with distinct properties from a single gene. *Eur. J. Biochem.* **262(2):** 575–585.
Morgenstern, B., Dress, A., and Werner, T. 1996. Multiple DNA and sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci* **93:** 12098–12103.
Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85:** 2444–2448.
Perier, R.C., Junier, T., and Bucher, P. 2000. The eukaryotic promoter database (EPD). *Nucl. Acids Res.* **28:** 302–303.
Quandt, K., Grote, K., and Werner, T. 1996. GenomeInspector: Basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. *Genomics* **33:** 301–304.
Reese, M.G., Hartzell G., Harris L.H., Ohler U., Abril J.F., and Lewis S. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10:** 483–501.
Scherf, M., Klingenhoff A., and Werner T. 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: A novel context analysis approach. *J. Mol. Biol.* **297(3):** 599–606.
Solovyev, V. and Salamov, A. 1997. The gene-finder computer tools for analysis of human and model organisms genome sequences. *Proc. Fifth Int. Conf. on Intelligent Syst. Mol. Biol. (ISMB97)* **5:** 294–302.
Werner, T. 1999. Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome* **10:** 168–175.

Received July 6, 2000; accepted in revised form December 29, 2000.