Assembly and Analysis of Extended Human Genomic Contig Regions

Eric C. Rouchka and David J. States

WUCS-99-10

March 8, 1998

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
Saint Louis, MO 63130-4899

Institute for Biomedical Computing
Washington University
700 S. Euclid Avenue
Saint Louis, MO 63110

ecr@ibc.wustl.edu
states@ibc.wustl.edu

Assembly and Analysis of Extended Human Genomic Contig Regions

Eric C. Rouchka and David J. States
Institute for Biomedical Computing
Washington University
700 South Euclid Avenue
St. Louis, MO 63110-1012, USA
Email: ecr@ibc.wustl.edu; states@ibc.wustl.edu

Abstract

The Human Genome Project (HGP) has led to the deposit of human genomic sequence in the form of sequenced clones into various databases such as the DNA Data Bank of Japan (DDBJ) (Tateno and Gojobori, 1997), the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (Stoesser, et. al., 1999), and GenBank (Benson, et. al., 1998). Many of these sequenced clones occur in regions where sequencing has taken place either within the same sequencing center or other centers throughout the world. The assembly of extended segments of genomic sequence by looking at overlapping end segments is desired and is currently available only in a limited sense from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/genome/seq/) and Oak Ridge National Laboratories' (ORNL) Genome Channel (http://compbio.ornl.gov/tools/channel/). We attempt to collate a definitive set of nonredundant extended segments of human genomic sequence by taking individual human entries in GenBank greater than 25 kilobases (kb) and extending them on either end. We address the several difficulties that arise when attempting to extend segments.

Keywords: Clone, contigs, genome analysis, annotation

Introduction

The U.S. Human Genome Project, coordinated by the United States Department of Energy (DOE) and the National Institutes of Health (NIH), began in 1990 as a 15 year venture to sequence the approximately three billion bases making up the human genome (Vaughan, 1996). As of October 1998, 180 million bases (6%) of the human genome has been sequenced (Collins, et. al., 1998).

Due to physical limitations of current sequencing and cloning techniques, the genome must be broken down into smaller portions in the range of 20 kilobases (kb) for plasmid clones to 250 kb for bacterial artificial chromosomes (BACs) and yeast artificial chromosomes (YACs) (Lodish, et. al., 1995). As the sequence data for each of these shorter regions becomes available, it would be helpful to connect them with adjacent overlapping regions previously sequenced. It is possible that overlapping sequences could originate from different sequencing centers. Since a complete sequence of each human chromosome is desired, a method to assemble these smaller sequences into larger contiguous regions (contigs) is constructed.

Methods

GenBank is used as the reference database for the human genomic DNA used in building the contigs. The results are based upon release 110.0, which includes sequences submitted to GenBank up until December 5, 1998. The GenBank primate division is used in order to create stable human contigs. In release 110.0, this is divided into gbpri1, gbpri2, and gbpri3. Table II shows a breakdown of the sequences in the primates division by sequence size.

Some of the genome sequencing centers incorporate neighboring clone information into their GenBank entries. Table I shows some examples of how this data is entered into the comments section. Use of this information could help in the creation of genome contigs. However, as Table I indicates, this data is not standardized among the sequencing centers. The data is entered by hand in a manner that is easy for a human to read, but not easily parsed by a computer. The overlap between two clones, if given, is present only in a positional manner. An alignment between two overlapping clones is not given.

We create most of the contigs using an automated procedure highlighted in Figure 1. The first step is to retrieve human sequences from GenBank which are greater than 25 kb in length. After these sequences are retrieved their ends are searched against the primate division of GenBank for overlapping regions at least 70 base pairs (bp) long, and at least 98% identical. These searches are performed using wu2blastn version 2.0 (Gish, 1994-1997), the Washington University version of BLAST (Altschul, et. al., 1990) with gaps for nucleic acid sequences.

Contigs can be extended by looking for blast hits to their ends. When overlapping clones are found, they are merged together into a contig based on the blast alignment. Discrepancies in the alignment resulting from gaps and mismatches are marked by the character N in the contig. After a set of contigs has been assembled, they are compared against contigs found at the NCBI and ORNL

Sequencing Center	GenBank Accession	Overlapping Information In COMMENT section
Sanger Centre	Z99715	The true right end of clone 1114G22 is at 104. The true left end of clone 262D12 is at 51983.
University of Washington Genome Sequencing Center	AC004398	Overlapping Sequences: 5': UWGC: g1248a010 (Accession: AC004107) 3': UWGC: g1248a139
Whitehead Institute for Biomedical Research	AC005303	Only 90.0 kilobases from the middle of this clone are being submitted. The remainder overlaps either accession AC003664 (WICGR project L281) or accession AC005277 (WICGR project L351).
Washington University Genome Sequencing Center	AC002378	NEIGHBORING SEQUENCE INFORMATION: The clone being sequenced to the left is BK085E05; the clone being sequenced to the right is DJ102K02. Actual start of this clone is at base position 1 of DJ438O4.
Baylor College of Medicine	AC002523	Begining of sequence overlaps with AF007262, end of sequence overlaps with AF011889. (*Note that Beginning is misspelled here*)

Table I: Overlapping clone information. The third column contains examples of overlapping clone information contained within the COMMENT section of the GenBank reports for the GenBank entries located in the second column. The overlapping clone information is typical for the sequencing centers shown in the first column.

web sites. Any differences are examined in more detail. In some cases, the restrictions need to be relaxed for automatic assembly to occur. Other contigs need to be assembled by hand in order to create the overlapping region. This is due to reported overlapping highest scoring pairs (Altschul, et. al., 1990) or other mismatches. Since the volume of sequencing data is growing exponentially, these steps are largely automated using Perl scripts.

Difficulties

There are several difficulties with trying to find overlapping end segments. One problem is that clones may not overlap with 100% identity due sequencing errors and polymorphisms. The Perl scripts are written in such a manner as to allow overlapping sequences greater than 98% identical. This allows the possibility that some overlaps might be missed. Most overlapping segments should be detected, however, since polymorphisms occur in the population at a rate of 7/1000 (Taillon-Miller, et. al., 1998), and acceptable sequencing error rates are 1/10000 (Collins, et. al., 1998).

Another difficulty is that the end of a sequence may contain repetitive elements. Prime examples of this are Alus and LINEs. In these cases, blast will produce multiple hits to otherwise unrelated sequences. It becomes

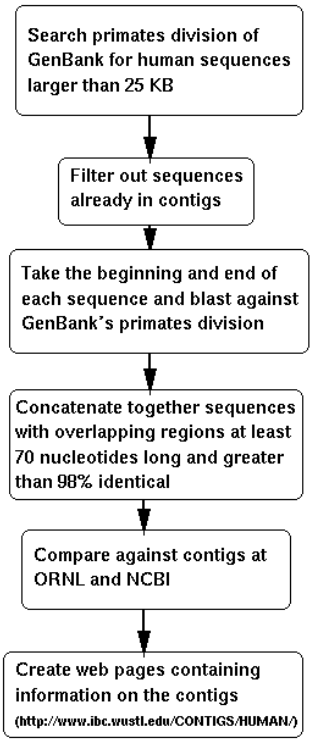

Figure 1: Contig creation flowchart. This figure indicates the steps that are followed in creating the human genomic contigs from GenBank entries.

hard to determine whether or not two sequences should be assembled into a contig when the overlap between them occurs in these repeat regions. Examples of such sequences are GenBank accession AC004021, AC004202, and AC004186.

The length of the overlap also varies greatly. Some sequencing centers such as Washington University Genome Sequencing Center (WUGSC) and Sanger Centre generally have a relatively constant sequence overlap length for known overlapping sequences. (In the case for WUGSC it is 200 bp; for Sanger Centre it is 100 bp.) For the assembled contigs, the size ranges from 0 base pair overlaps from the Japan Science and Technology Corporation efforts on chromosome 21 to a 82,766 base overlap between GenBank accession HS326L12 and HS232G24 from Sanger on chromosome X. Note that those sequences with less than a 70 base pair overlap are hand assembled. The GenBank entries for these sequences have been used to aid in the detection and assembly of these contigs. For the shorter overlapping segments, running blast to find the alignment between two sequences takes a matter of seconds, but for larger regions, the time spent to find the alignment can take hours.

Results

GenBank release 110.0 contains 2,644 human genomic sequences greater than 25 kb in length. Table II indicates the breakdown of these clones. As of January 8, 1999, we have assembled a total of 1608 contigs. These contigs cover a total of 228,773,482 bases. Note that there are more clones in the assembled contigs than entries in GenBank greater than 25 kb due to the fact that several contigs contain clones shorter than 25 kb.

Table III indicates the breakdown of the contigs by their size. This shows that most of the contigs are comprised of either one or two clones. There are three examples that contain 20 or more clones, including a 2,138,766 base region on chromosome 7q31.3 which is comprised of 65 clones. This effort part of the University of Washington Genome Center's chromosome 7 sequencing project (Iadonato, et. al., 1996). Most of the contigs lie within the 0-300 kb range. There are six contigs larger than 1 Mb in length. The largest of these is the contig described above.

Sequence Size (in nucleotides)	Number of GenBank entries
> 200,000	72
150,000-199,999	324
100,000-149,999	681
75,000-99,999	340
50,000-74,999	169
25,000 -49,999	1058
TOTAL > 25,000	2,644

Table II: Size of primate GenBank entries. This table indicates the number of sequences in the primate divisions (gbpri1, gbpri2, and gbpri3) of GenBank release 110.0.

Contig Size (in clones)	Number of contigs
1	1154
2	264
3	82
4	33
5	24
6	15
7	7
8	9
9	4
10	6
11-20	7
20+	3

Contig Size (in kilobases)	Number of contigs
0-50	345
50-100	287
100-150	449
150-200	251
200-300	169
300-400	63
400-500	13
500-1000	25
1000+	6

Table III: Size of generated contigs. The left-hand portion of this table indicates the number of contigs falling within a size range where the size is the number of GenBank entries which are concatenated together to produce them. The right-hand portion of the table reports the number of contigs falling within a certain size range, where the size is based on the number of nucleotides in the contig.

The breakdown by chromosome is presented in Table IV. A comparison between the contigs we have assembled and the ones collated at NCBI is presented. According to this data, the contigs cover about 7% of the human genome through GenBank release 110.0, consistent with the current status of the Human Genome Project (Collins, et. al., 1998).

By examining Table IV, is can be seen that chromosome 7 and chromosome X have been the most heavily sequenced chromosomes. Chromosome 22 is closer to being completely sequenced, due to its shorter length.

Discrepancies

There have been at least three different situations where clones from different chromosomes overlap to form chimeric contigs which have been uncovered as a result of assembling human genomic contigs. Two of these examples seem to be related to errors in the GenBank

CHROMOSOME		IBC Contigs (1/9/1999)		NCBI Contigs (12/2/1998)	
Num.	Size (MB)	Size (MB)	Percent Covered	Size (MB)	Percent Covered
1	263	7.4	2.8	6.1	2.3
2	255	1.1	0.4	0.6	0.2
3	214	1.5	0.7	1.4	0.6
4	203	5.3	2.6	4.1	2.0
5	194	11.1	5.7	10.1	5.2
6	183	18.3	10.0	13.9	7.6
7	171	37.7	22.0	32.4	18.9
8	155	1.7	1.1	1.3	0.9
9	145	2.1	1.4	1.5	1.0
10	144	1.6	1.1	0.1	0.1
11	144	5.5	3.8	3.9	2.7
12	143	7.2	5.0	5.0	3.5
13	98	1.9	1.9	1.8	1.8
14	93	2.3	2.5	1.5	1.6
15	98	1.8	1.8	1.3	1.4
16	93	15.3	16.5	13.1	14.1
17	89	21.7	24.4	13.3	15.0
18	85	0.2	0.2	0.1	0.1
19	67	11.0	16.4	9.7	14.5
20	72	2.3	3.2	1.5	2.0
21	39	10.4	26.7	9.3	23.8
22	43	16.2	37.7	13.5	31.5
X	164	42.5	25.9	36.5	22.3
Y	59	1.4	2.4	0.8	1.4
Unknown	N/A	1.3	N/A	N/A	N/A
TOTALS	3,214	228.8	7.1	183.0	5.7

Table IV: This table represents the results as of January 8, 1999. The expected chromosome size in the second column is taken from the NCBI page. The contig sizes in column 5 are taken from the NCBI page: **http://www.ncbi.nlm.nih.gov/genome/seq**. The IBC contigs are the contigs that we have assembled and are available at the location **http://www.ibc.wustl.edu/contigs/HUMAN/**. All of the data in this table represents nonredundant sequence data. Note that the NCBI site has been inaccessible since December 1998.

entries. The third example appears to have support for being caused by a real biological event.

Mislabeled GenBank Entry

In one case, two overlapping clones are observed that were annotated as originating from two separate chromosomes. The clones are both from the University of Texas Southwestern Medical Center (UTSW) where sequencing on chromosomes 15 and 11 are active projects. The first clone has GenBank accession AC002426 and the second has accession AC002468. AC002426 is 118 kb in length and AC002468 is 116 kb in length. The end of AC002426 overlaps the beginning of AC002468 by 41kb with 99.98% identity. AC002426 is annotated as originating from 15q26, while AC002468 was originally annotated as originating from chromosome 11. UTSW acknowledged the misannotation of AC002468 and has since updated the GenBank entry.

A contig from two different chromosomes

A second interesting region occurs when three separate clones are involved in creating a chimeric contig from different chromosomes. The three entries are Z99571 (125 kb) , Z82216 (142 kb), and AC003001 (102 kb). Z99571 and Z82216 are sequenced at Sanger Centre and are mapped to Xq21.1. AC003001 is sequenced at Whitehead Institute/MIT Center for Genome Research and is mapped to chromosome 17. A contig can be constructed since the end of Z99571 overlaps the beginning of AC003001 by 20 kb without any mismatches or indels. The end of AC003001 overlaps the beginning of Z82216 at 99.990% identity. There are three indels and one mismatch event occurring in this 41kb overlap. The high identity overlaps for this region indicate that it is most likely that either both of the Sanger entries are misannotated, or the MIT entry is misannotated.

Two overlapping clones from different chromosomes

Another interesting region occurs between two overlapping clones originating from two separate chromosomes. The first entry is GenBank accession AL021921 and the second entry is GenBank accession U95738. The 135 kb AL021921 is sequenced by Sanger Centre and is annotated as 1p36.13. The 171 kb entry U95738 is sequenced by The Institute for Genome Research (TIGR) and is annotated as 16p13.11. According to the blast hits, AL021921 lies completely within U95738 with 100 mismatches, 74 of which are transitions (A⇔G; C⇔T) and 26 are transversions. There are also 22 gaps composed of 123 indel events. At random, it is expected to have twice as many transversions as transitions. However, in this case, there are almost three times as many transitions as transversions. FISH results from the California Institute of Technology indicate that there is a duplication event occurring between chromosomes 1 and 16. This data is given in Figure 2.

Analysis of Contig Regions

Once extended human genomic contigs have been assembled, they can be analyzed. Among the plans for analysis include looking for a correlation between microsatellite repeats and nucleotide composition, statistical analysis between different regions, and looking at evolutionary distance measures. Another practical use for these extended segments involves using them to position end sequenced clones for the purpose of sequence validation and location of polymorphisms.

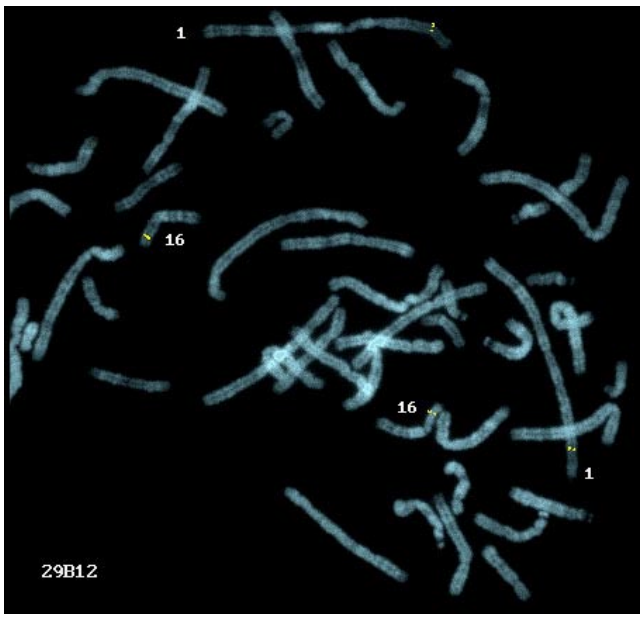

Figure 2: FISH results for GenBank accession U95738. This FISH image indicates a homology between chromosomes 1 and 16. (Image courtesy of California Institute of Technology) (http://www.tree.caltech.edu/pictures.fish-29B12.jpg)

In addition to the data presented in Table IV, there is a total of 7,693,559 bases contained within 134 sequenced clones that overlaps contigs already assembled. This data will be compared with the assembled clones. Hopefully these extra sequences will lend some information into the distribution of single nucleotide polymorphisms and mutational hotspots.

Discussion

In an ideal situation, there would be only one way for the clone pieces of the genomic puzzle to fit together. As the example of two overlapping clones from different chromosomes shows above, this is not always the case.

Due to natural selection, the human genome tends to retain those parts of the genome that are useful and reuses them. As a result, there is nonrandomness associated with genomic data. This makes it difficult to verify whether or not two clones do indeed exist in a contig or they just happen to have some similarities in their ends. This will become a more prevalent problem as more and more data becomes available through the Human Genome Project.

All of the contigs in the NCBI Human Genome Sequencing pages with clones are greater than 25 kb in length are contained within the assembled clones. In addition, we have approximately 14 Mb more data than NCBI. The Genome Channel provides a graphical Java interface which is nice in trying to locate the placement of contigs within chromosomes. ORNL also offers extensive annotation on the contigs. However, the Genome Channel only contains about 180 Mb of total sequence data as of January 22, 1999, which is approximately 80% of the data that we have made available.

Summary

Automated assembly of human genomic contigs is a useful endeavor. There are some issues requiring human intervention in order to maintain the integrity of the contigs being built. Once the contigs are assembled, analysis can proceed into understanding a more large scale or macro view of the differences in composition across the human genome. Due to the expected exponential growth of data available in the genomic databases, it is becoming imperative that procedures become automated to create and annotate these large sequences. It is equally important to be able to determine which sequences are redundant and which offer novel information.

Acknowledgements

We wish to thank Brendan Loftus of the TIGR Center and Andrew King of the Sanger Center for assistance in reviewing clone origins and FISH data.

We also wish to thank Thomas Blackwell, David Maffitt, Volker Nowotny, and David Politte for many hours of thoughtful discussion and critique.

This work is funded in part by a grant from the United States Department of Energy grant ER61910 000261 with David States as the primary investigator.

All of the contigs created using our methods are available at the URL: http://www.ibc.wustl.edu/CONTIGS/HUMAN/

References

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., (1998) "GenBank." *Nucleic Acids Research*, **26**(1):1-7.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., (1990) "Basic Local Alignment Search Tool." *Journal of Molecular Biology*, **215**, 403-410.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., (1998) "New Goals for the U.S. Human Genome Project: 1998-2003." *Science*, **282**(5389): 682-689.

Gish, W., (1994-1997). unpublished.

Iadonato, S.P., Yu, J., Wong, G.K.-S., Magness, C.L., Green, E.D., Green, P., Olson, M.V., (1996) "Large-scale MCD Mapping and Sequencing of Human Chromosome 7." unpublished.

Lodish, H., Baltimore, D., Berk, A., Zipursky, S.L., Matsudaria, P., Darnell, J. (1995). *Molecular Cell Biology*. New York: Scientific American Books.

Stoesser, G., Tuli, M.A., Lopez, R., Sterk, P., (1999) "The EMBL Nucleotide Sequence Database." *Nucleic Acids Research*, **27**(1):18-24.

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., Kwok, P.Y., (1998) "Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms." *Genome Research* **8**(7):748-754.

Tateno, Y., Gojobori, T., (1997) "DNA Data Bank of Japan in the age of information biology." *Nucleic Acids Research*, **25**(1):14-7.

Vaughan, D. (ed.), (1996) "To Know Ourselves" http://www.ornl.gov/TechResources/Human_Genome/tko/