An Example of Information Management in Biology: Qualitative Data Economizing Theory Applied to the Human Genome Project Databases

Iraj Daizadeh
Research Informatics, Amgen, Inc., One Amgen Center Drive, Thousand Oaks, CA 91320-1799.
E-mail: IrajDaizadeh@yahoo.com

Ironically, although much work has been done on elucidating algorithms for enabling scientists to efficiently retrieve relevant information from the glut of data derived from the efforts of the Human Genome Project and other similar projects, little has been performed on optimizing the levels of *data economy* across databases. One technique to qualify the degree of data economization is that constructed by Boisot. Boisot's Information Space (I-Space) takes into account the degree to which data are written (codification), the degree to which the data can be understood (abstraction), and the degree to which the data are effectively communicated to an audience (diffusion). A data system is said to be more data economical if it is relatively *high* in these dimensions. Application of the approach to entries in two popular, publicly available biological data repositories, the Protein DataBank (PDB) and GenBank, leads to the recommendation that PDB increases its level of abstraction through establishing a larger set of detailed keywords, diffusion through constructing hyperlinks to other databases, and codification through constructing additional subsections. With these recommendations in place, PDB would achieve the greater data economies currently enjoyed by GenBank. A discussion of the limitations of the approach is presented.

Introduction

Due to the results of the Human Genome Project and other data intensive technological advances, knowledge management is now playing a critical role in today's science (see, e.g., Politz, van Driel, Sauer, & Pombo, 2003). Unfortunately, due to differences in codification, abstraction, and diffusion, there are differences in the degree of data economies afforded by these various approaches. Here, I analyze two such scientific databases, the Protein Data Bank (PDB) and GenBank, through the lens of Boisot's Information Space model (herein, the Boisot Cube or I-Space) to elucidate how the structure of the information presented within these databases proffers varying amounts of economic value.

For Boisot (1999), knowledge assets arose from the firm's attempts to economize data processing—the transformation of raw data into meaningful information that can be used by the firm. To economize data, Boisot introduced a single integrated conceptual framework—termed the *I-Space*—which takes into account: the degree to which data are written, the degree to which the data can be understood, and the degree to which the data are effectively communicated to the largest audience (Boisot, 1999). Thus, the three dimensions for constructing the I-Space are codification, abstraction, and diffusion, respectively. The higher a system is in all three of these dimensions, the more data economical it will be (Biosot, 1999). The I-Space construct is shown in Figure 1.

Points within the I-Space detail the degree of codification, abstraction, and diffusion of a particular unit of data or information, and flows within the cube can be interpreted either sequentially or chronologically. The I-Space model has been used to investigate theories of learning and culture (Boisot, 1995) and organizational growth (Boisot, 1999), among other applications (Boisot & Child, 1996).

To the author's knowledge, this is the first application of the I-Space model to investigate data economization within biological databases. In the next section, I describe how the I-Space model can be applied to compare two of the most popular and largest publicly available biological data repositories. This article concludes with a discussion of practical suggestions that the PDB authors may use to optimize data economies currently enjoyed by those of GenBank. This brief communication concludes with a discussion of the limitations of the approach, and steps for further development.

Received September 3, 2004; revised November 9, 2004; accepted January 10, 2005

• Published online 11 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20270

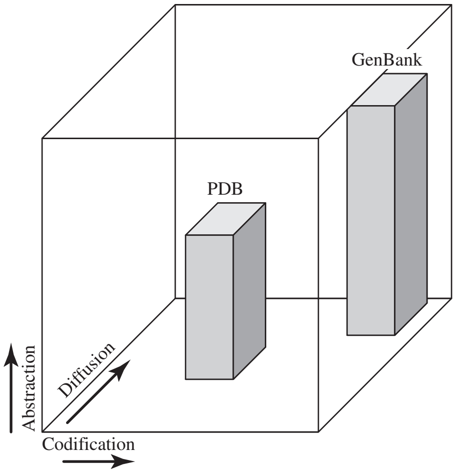


FIG. 1. The PDB and GenBank mapped within the Boisot Cube; definitions of the three unit vectors are described in the text.

Method

The following definitions of Boisot's unit vectors were used as the basis for constructing the positioning of the PDB and GenBank data elements—namely, particular data entries—within the I-Space construct. Below I provide an overview of the dimensions used to construct the I-Space model; readers are referred to Boisot's works in 1999 and 1986 for further information.

- Codification: Boisot defines codification as the degree to which the knowledge is written into transmittable form, for example, laboratory notebooks, books, patents, and so on. A recipe or a patent may be considered a well-codified document because the practitioner who has followed the required steps can reproduce—within some small level of uncertainty—the results of the experiment. In the limit, codification "then allows a task to be performed entirely by machine without human intervention (Boisot, 1999, p. 47)." On the other hand, the precise process of making a car, for example, following Toyota's just-in-time model, would be very difficult to write down in a series of simple steps, and is thus poorly codified. Boisot has defined tacit knowledge—knowledge that is inarticulate, complex, and noncodified—in this limit (Boisot, 1999; Polanyi, 1958). Thus, in Boisot's formalism, tacit knowledge includes existential, endemic, and experiential knowledge as well (see Doz, Santos, & Williamson, 2001).
- Abstraction: Abstraction corresponds to the degree to which the (economized) data can be understood. To illustrate the extremes of this dimension, one can consider elementary ideas in physics versus those in biology. Newton's equations can be used to track the location of any mass moving classically in physical three-dimensional space. These equations are sufficiently general that any path of a traveling particle—irrespective of size and behaving within the classical and nonrelativistic regime—may be mathematically traced in three-dimensional space. On the other hand, knowledge of glucose-6-phosphatase and its use in converting glucose-6-phosphate into glucose is only one reaction within a large and complex metabolic pathway. In Boisot's formalism, this latter extreme may be considered "predominantly perceptual and local" and thus definable as a single concrete manifestation or single instantiation of knowledge, while abstraction illustrated within Newton's formalism supports an extendable conceptual knowledge capable of extensibility (Brinklow, 2004). Thus, the greater the degree of abstraction, as defined by Boisot, the more generally applicable the outcome, and thus the greater efficiency in economizing data (Boisot, 1999).
- Diffusion: The degree of diffusibility corresponds to the "proportion of a given population of data-processing agents (e.g., individuals, firms, industries, countries) that can be reached with information operating at different degrees of codification and abstraction" (Boisot, 1999, p. 52). As an example of highly diffuse data or information, a recipe can be easily distributed in an e-mail to thousands of individuals, irrespective of cooking experience, with the exact methodology for cooking a pie. In the other extreme, esoteric or inarticulate knowledge, such as wants and desires, are difficult to diffuse to a given population, because such elements of vernacular elicit different meanings to different individuals within a population.

The application of the Boisot method for examining data elements followed directly from the above definitions. Assuming that both GenBank and PDB have standard formats for their database entries, two data entries within the PDB and GenBank databases, as shown in Figures 2 through 4, were randomly extracted and analyzed within the qualitative model, based on the above definitions. The content search was performed manually, where visually the data entries were probed for differences and similarities. The results of these comparisons were then categorized based on the I-space definitions presented above. A discussion of the approach, including caveats of the method, appears below. From the analysis, Figure 1 was constructed.

Discussion and Conclusions

As of April 22, 2004, 28 million sequences and 28 billion base pairs (bits) resulting from the various sequencing projects, including the human genome project, were housed within GenBank; and the PDB contained the three-dimensional structures of 20,747 biological molecules. Figure 2 presents an overview of these databanks; typical entries for GenBank and PDB are presented in Figures 3 and 4, respectively.

By investigating Figures 3 and 4, and looking for differences in the degree of abstraction, codification, and diffusion, we find the following results. GenBank has *keywords* specifying each particular section of an entry. These keywords seem to correspond to the nature of the contents under that section. For example, *organism* is associated with *Mus musculus* (the mouse), along with the taxonomic linkages from which this species is derived. Although the PDB entry does have a keywords, these are not indicative of the

DOI: 10.1002/asi

"GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research* 2002 Jan 1;30(1):17-20). There are approximately 22,617,000,000 bases in 18,197,000 sequence records as of August 2002 (see GenBank growth statistics). As an example, you may view the record for a *Saccharomyces cerevisiae* gene. The complete release notes for the current version of GenBank are available. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis."

"The Protein Data Bank (PDB) is the single international repository for public data on the 3-dimensional structures of biological macromolecules. The contents are primarily experimental data derived from X-ray crystallography and NMR experiments. The primary goals of this resource are:

- To enable you to locate structures of interest;
- To perform simple analyses on one or more structures;
- To act as a portal to additional information available on the Internet;
- To enable you to download information on a structure, notably the Cartesian atomic coordinates, for further analysis.

The database is constantly updated as new structures are deposited by the international scientific community."

FIG. 2. Brief overview of the GenBank and PDB Databases. (The first vignette was extracted from the GenBank Web page: http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html/. The second was extracted from the PDB Web page: http://www.rcsb.org/pdb/help-general.html#What.)

contents of the section. For example, *remark* may contain information concerning references, or a description of the sequence features (see, *e.g.,* PDB lines: 1CBN 64 and 1CBN 54). PDB attempts to separate contents through the use of secondary keywords, such as *remark 1 titl*; however, we are unsure if this corresponds to reference 1 (see line: 1CBN 16), or reference 2 (see line: 1CBN 22). Thus, GenBank entries are more abstract than PDB entries—in Boisot's verbiage, the keyword *remark* is overly concrete and is not generally applicable.

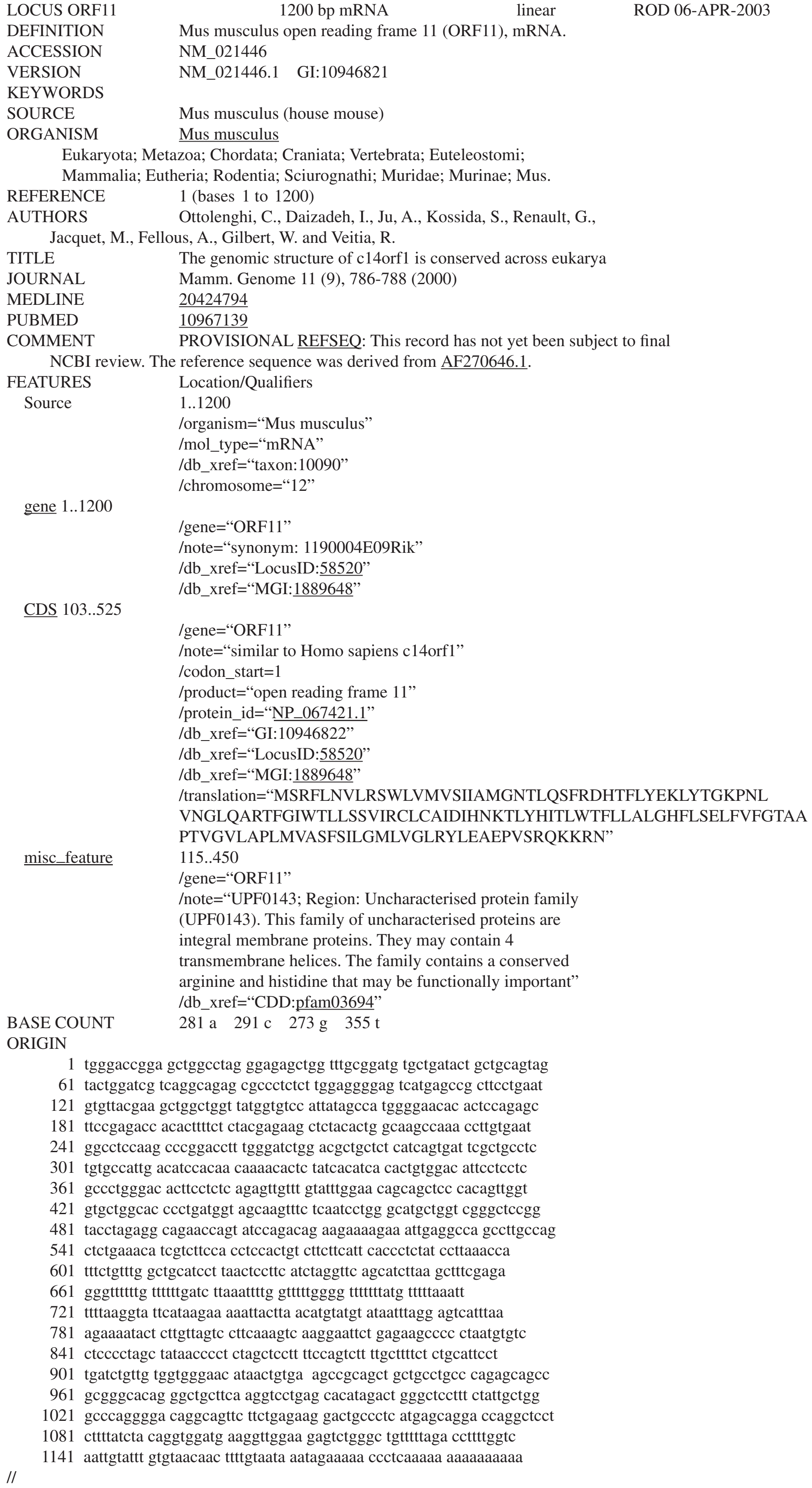
Moreover, the GenBank format also allows for the greatest degree of *diffusion* with respect to that of the PDB. The reason for this lies in the fact that various elements within the GenBank entry are underlined—these are hyperlinks to entries within other databases, even outside of the country in which GenBank is located. Thus, users can obtain more detailed information. PDB could have had links, even to GenBank, for example; it does not have this facility at all.

Finally, each GenBank entry presents information beyond that of the simple sequence, such as taxonomical, literature, and other interesting information. The PDB entry—effectively—has only the 3-dimensional coordinates of each atom within a particular structure. PDB could have been more codified, for example, by including taxonomical information. Figure 1 summarizes the key findings of this paper.

In summary, it is recommended that the PDB increase its levels of abstraction, through establishing a larger set of detailed keywords; diffusion, through constructing hyperlinks with other databases; and codification, through constructing more subsections. With these recommendations in place, PDB would achieve the greater data economies currently enjoyed by GenBank.

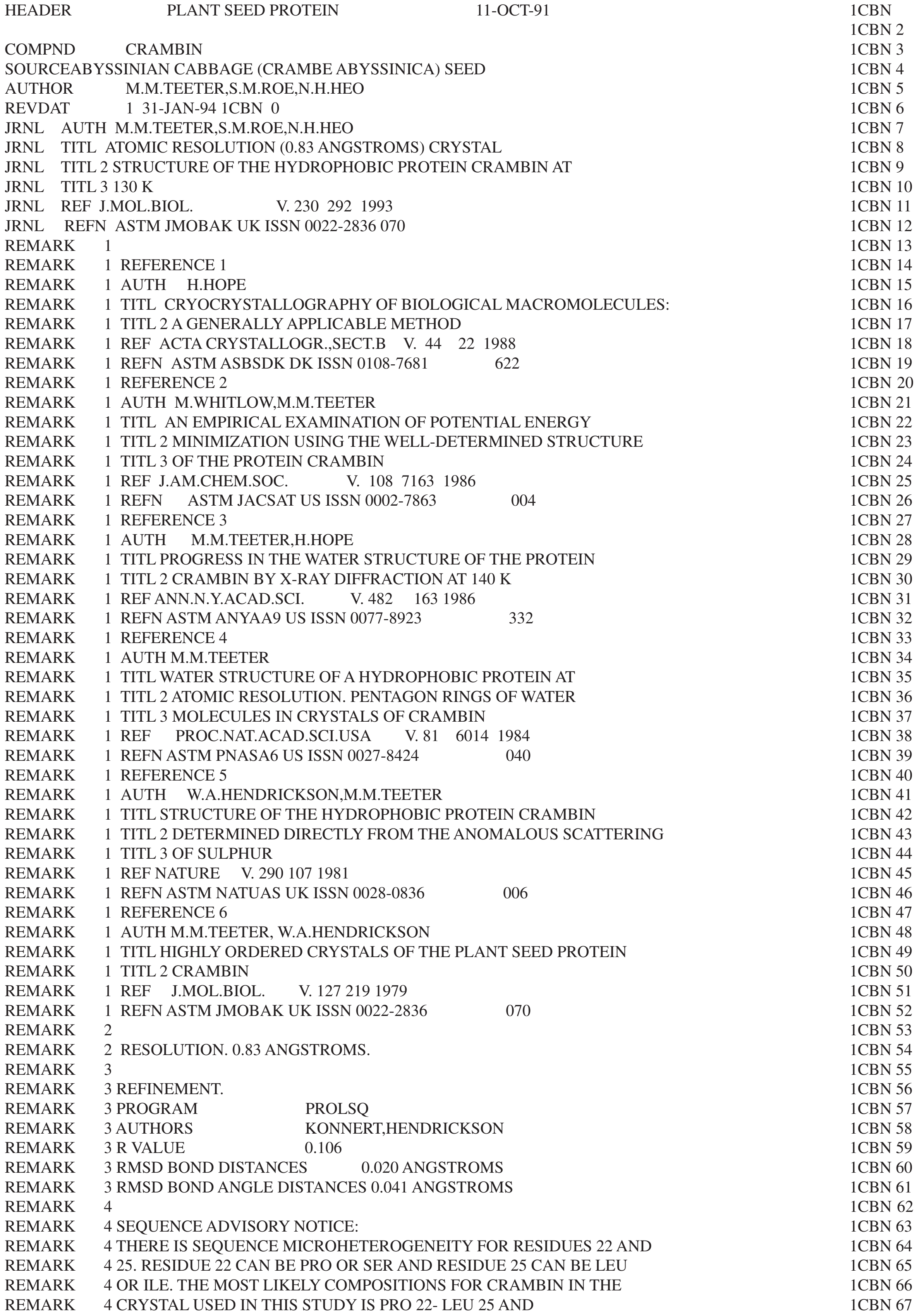
In conclusion, this application of the I-Space model shows how a descriptive tool can qualify, but not quantify, data economies for various knowledge management approaches. There are several caveats to the approach taken here. Indeed, Boisot states the dilemma clearly: "knowledge management theories often generate theories that are too general or abstract to be easily testable;" he has recently published a note on simulation modeling of the I-Space (see, e.g., Boisot, Canals, & MacMillan, 2004). There lies the potential for *heuristic* investigation leading to the general positioning of data within the I-Space. One such approach may include a simple random comparison like that performed here, or more sophisticated survey techniques applied to a population of users. Statistical inferences from such surveys may yield insight into the abstraction and diffusion axes, whereas codification may be investigated through a computational content search. The author is currently investigating the feasibility of these other methods, and welcomes any collaboration.

It is our hope, with this brief application, to attract information theorists to this line of inquiry. Extending such potentially promising approaches may yield insight useful for optimizing the information-gathering approaches that have thus far hampered full appreciation of large data banks, such as those commonly found in the large-scale genomic sequencing projects.



```
LOCUS ORF11                  1200 bp mRNA            linear        ROD 06-APR-2003
DEFINITION            Mus musculus open reading frame 11 (ORF11), mRNA.
ACCESSION             NM_021446
VERSION               NM_021446.1    GI:10946821
KEYWORDS
SOURCE                Mus musculus (house mouse)
ORGANISM              Mus musculus
      Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
      Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
REFERENCE             1 (bases  1 to  1200)
AUTHORS               Ottolenghi, C., Daizadeh, I., Ju, A., Kossida, S., Renault, G.,
    Jacquet, M., Fellous, A., Gilbert, W. and Veitia, R.
TITLE                 The genomic structure of c14orf1 is conserved across eukarya
JOURNAL               Mamm. Genome 11 (9), 786-788 (2000)
MEDLINE               20424794
PUBMED                10967139
COMMENT               PROVISIONAL REFSEQ: This record has not yet been subject to final
    NCBI review. The reference sequence was derived from AF270646.1.
FEATURES              Location/Qualifiers
  Source              1..1200
                      /organism="Mus musculus"
                      /mol_type="mRNA"
                      /db_xref="taxon:10090"
                      /chromosome="12"
  gene 1..1200
                      /gene="ORF11"
                      /note="synonym: 1190004E09Rik"
                      /db_xref="LocusID:58520"
                      /db_xref="MGI:1889648"
  CDS 103..525
                      /gene="ORF11"
                      /note="similar to Homo sapiens c14orf1"
                      /codon_start=1
                      /product="open reading frame 11"
                      /protein_id="NP_067421.1"
                      /db_xref="GI:10946822"
                      /db_xref="LocusID:58520"
                      /db_xref="MGI:1889648"
                      /translation="MSRFLNVLRSWLVMVSIIAMGNTLQSFRDHTFLYEKLYTGKPNL
                      VNGLQARTFGIWTLLSSVIRCLCAIDIHNKTLYHITLWTFLLALGHFLSELFVFGTAA
                      PTVGVLAPLMVASFSILGMLVGLRYLEAEPVSRQKKRN"
  misc_feature        115..450
                      /gene="ORF11"
                      /note="UPF0143; Region: Uncharacterised protein family
                      (UPF0143). This family of uncharacterised proteins are
                      integral membrane proteins. They may contain 4
                      transmembrane helices. The family contains a conserved
                      arginine and histidine that may be functionally important"
                      /db_xref="CDD:pfam03694"
BASE COUNT            281 a    291 c    273 g    355 t
ORIGIN
         1  tgggaccgga gctggcctag ggagagctgg tttgcggatg tgctgatact gctgcagtag
        61  tactggatcg tcaggcagag cgccctctct tggaggggag tcatgagccg cttcctgaat
       121  gtgttacgaa gctggctggt tatggtgtcc attatagcca tggggaacac actccagagc
       181  ttccgagacc acacttttct ctacgagaag ctctacactg gcaagccaaa ccttgtgaat
       241  ggcctccaag cccggacctt tgggatctgg acgctgctct catcagtgat tcgctgcctc
       301  tgtgccattg acatccacaa caaaacactc tatcacatca cactgtggac attcctcctc
       361  gccctgggac acttcctctc agagttgttt gtatttggaa cagcagctcc cacagttggt
       421  gtgctggcac ccctgatggt agcaagtttc tcaatcctgg gcatgctggt cgggctccgg
       481  tacctagagg cagaaccagt atccagacag aagaaaagaa attgaggcca gccttgccag
       541  ctctgaaaca tcgtcttcca cctccactgt cttcttcatt caccctctat ccttaaacca
       601  tttctgtttg gctgcatcct taactccttc atctaggttc agcatcttaa gctttcgaga
       661  gggtttttg tttttgatc ttaaattttg gtttttgggg tttttttatg tttttaaatt
       721  ttttaaggta ttcataagaa aaattactta acatgtatgt ataatttagg agtcatttaa
       781  agaaaatact cttgttagtc cttcaaagtc aaggaattct gagaagcccc ctaatgtgtc
       841  ctcccctagc tataacccct ctagctcctt ttccagtctt ttgcttttct ctgcattcct
       901  tgatctgttg tggtgggaac ataactgtga  agccgcagct gctgcctgcc cagagcagcc
       961  gcgggcacag ggctgcttca aggtcctgag cacatagact gggctccttt ctattgctgg
      1021  gcccagggga caggcagttc ttctgagaag gactgccctc atgagcagga ccaggctcct
      1081  cttttatcta caggtggatg aaggttggaa gagtctgggc tgtttttaga ccttttggtc
      1141  aattgtattt gtgtaacaac ttttgtaata aatagaaaaa ccctcaaaaa aaaaaaaaaa
//
```

FIG. 3. A typical entry within GenBank, culled from http://www.ncbi.nlm.nih.gov. Notice the KEYWORDS on the left in CAPITAL letters; and links to other parts of the database are underlined. For further information concerning the KEYWORDS and contents, the interested reader is forwarded to http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html for a description of a sample record.



```
HEADER          PLANT SEED PROTEIN                 11-OCT-91                              1CBN
                                                                                          1CBN 2
COMPND        CRAMBIN                                                                     1CBN 3
SOURCEABYSSINIAN CABBAGE (CRAMBE ABYSSINICA) SEED                                         1CBN 4
AUTHOR        M.M.TEETER,S.M.ROE,N.H.HEO                                                  1CBN 5
REVDAT        1  31-JAN-94 1CBN  0                                                        1CBN 6
JRNL    AUTH  M.M.TEETER,S.M.ROE,N.H.HEO                                                  1CBN 7
JRNL    TITL  ATOMIC RESOLUTION (0.83 ANGSTROMS) CRYSTAL                                  1CBN 8
JRNL    TITL 2 STRUCTURE OF THE HYDROPHOBIC PROTEIN CRAMBIN AT                            1CBN 9
JRNL    TITL 3 130 K                                                                      1CBN 10
JRNL    REF  J.MOL.BIOL.              V. 230  292  1993                                   1CBN 11
JRNL     REFN  ASTM JMOBAK UK ISSN 0022-2836 070                                          1CBN 12
REMARK      1                                                                             1CBN 13
REMARK      1  REFERENCE 1                                                                1CBN 14
REMARK      1  AUTH     H.HOPE                                                            1CBN 15
REMARK      1  TITL  CRYOCRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES:                    1CBN 16
REMARK      1  TITL 2 A GENERALLY APPLICABLE METHOD                                       1CBN 17
REMARK      1  REF  ACTA CRYSTALLOGR.,SECT.B    V.  44    22  1988                        1CBN 18
REMARK      1  REFN  ASTM ASBSDK DK ISSN 0108-7681            622                         1CBN 19
REMARK      1  REFERENCE 2                                                                1CBN 20
REMARK      1  AUTH  M.WHITLOW,M.M.TEETER                                                 1CBN 21
REMARK      1  TITL  AN EMPIRICAL EXAMINATION OF POTENTIAL ENERGY                         1CBN 22
REMARK      1  TITL 2 MINIMIZATION USING THE WELL-DETERMINED STRUCTURE                    1CBN 23
REMARK      1  TITL 3 OF THE PROTEIN CRAMBIN                                              1CBN 24
REMARK      1  REF  J.AM.CHEM.SOC.          V.  108  7163  1986                           1CBN 25
REMARK      1  REFN      ASTM JACSAT US ISSN 0002-7863              004                   1CBN 26
REMARK      1  REFERENCE 3                                                                1CBN 27
REMARK      1  AUTH      M.M.TEETER,H.HOPE                                                1CBN 28
REMARK      1  TITL PROGRESS IN THE WATER STRUCTURE OF THE PROTEIN                        1CBN 29
REMARK      1  TITL 2 CRAMBIN BY X-RAY DIFFRACTION AT 140 K                               1CBN 30
REMARK      1  REF ANN.N.Y.ACAD.SCI.         V. 482     163 1986                          1CBN 31
REMARK      1  REFN ASTM ANYAA9 US ISSN 0077-8923                   332                   1CBN 32
REMARK      1  REFERENCE 4                                                                1CBN 33
REMARK      1  AUTH M.M.TEETER                                                            1CBN 34
REMARK      1  TITL WATER STRUCTURE OF A HYDROPHOBIC PROTEIN AT                           1CBN 35
REMARK      1  TITL 2 ATOMIC RESOLUTION. PENTAGON RINGS OF WATER                          1CBN 36
REMARK      1  TITL 3 MOLECULES IN CRYSTALS OF CRAMBIN                                    1CBN 37
REMARK      1  REF      PROC.NAT.ACAD.SCI.USA        V. 81    6014  1984                  1CBN 38
REMARK      1  REFN ASTM PNASA6 US ISSN 0027-8424                   040                   1CBN 39
REMARK      1  REFERENCE 5                                                                1CBN 40
REMARK      1  AUTH     W.A.HENDRICKSON,M.M.TEETER                                        1CBN 41
REMARK      1  TITL STRUCTURE OF THE HYDROPHOBIC PROTEIN CRAMBIN                          1CBN 42
REMARK      1  TITL 2 DETERMINED DIRECTLY FROM THE ANOMALOUS SCATTERING                   1CBN 43
REMARK      1  TITL 3 OF SULPHUR                                                          1CBN 44
REMARK      1  REF NATURE     V. 290 107 1981                                             1CBN 45
REMARK      1  REFN ASTM NATUAS UK ISSN 0028-0836                006                      1CBN 46
REMARK      1  REFERENCE 6                                                                1CBN 47
REMARK      1  AUTH M.M.TEETER, W.A.HENDRICKSON                                           1CBN 48
REMARK      1  TITL HIGHLY ORDERED CRYSTALS OF THE PLANT SEED PROTEIN                     1CBN 49
REMARK      1  TITL 2 CRAMBIN                                                             1CBN 50
REMARK      1  REF     J.MOL.BIOL.        V. 127 219 1979                                 1CBN 51
REMARK      1  REFN ASTM JMOBAK UK ISSN 0022-2836                070                      1CBN 52
REMARK      2                                                                             1CBN 53
REMARK      2  RESOLUTION. 0.83 ANGSTROMS.                                                1CBN 54
REMARK      3                                                                             1CBN 55
REMARK      3 REFINEMENT.                                                                 1CBN 56
REMARK      3 PROGRAM                    PROLSQ                                           1CBN 57
REMARK      3 AUTHORS                    KONNERT,HENDRICKSON                              1CBN 58
REMARK      3 R VALUE                    0.106                                            1CBN 59
REMARK      3 RMSD BOND DISTANCES             0.020 ANGSTROMS                             1CBN 60
REMARK      3 RMSD BOND ANGLE DISTANCES 0.041 ANGSTROMS                                   1CBN 61
REMARK      4                                                                             1CBN 62
REMARK      4 SEQUENCE ADVISORY NOTICE:                                                   1CBN 63
REMARK      4 THERE IS SEQUENCE MICROHETEROGENEITY FOR RESIDUES 22 AND                    1CBN 64
REMARK      4 25. RESIDUE 22 CAN BE PRO OR SER AND RESIDUE 25 CAN BE LEU                  1CBN 65
REMARK      4 OR ILE. THE MOST LIKELY COMPOSITIONS FOR CRAMBIN IN THE                     1CBN 66
REMARK      4 CRYSTAL USED IN THIS STUDY IS PRO 22- LEU 25 AND                            1CBN 67
```

FIG. 4. A typical entry within the Protein Data Bank (PDB), culled from http://www.rcsb.org. Notice the KEYWORDS on the left in CAPITAL letters. For further information concerning the KEYWORDS and contents, the interested reader is forwarded to http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html for a description of the record. Some atomic coordinates have been removed to shorten the record whilst not altering the overall structure of the record.

DOI: 10.1002/asi

```
REMARK        4 SER 22- ILE 25. BECAUSE OF LIMITATIONS IN PROTEIN DATA                1CBN 68
REMARK        4 FORMAT, ONLY PRO 22 AND LEU 22 ARE SHOWN ON THE SEQRES                1CBN 69
REMARK        4 RECORDS BELOW. IN ADDITION RESIDUES SER 22 AND ILE 25 ARE             1CBN 70
REMARK        4 PRESENTED AS RESIDUES SER 22B AND ILE 25B ON THE ATOM                 1CBN 71
REMARK        4 RECORDS BELOW, IMMEDIATELY FOLLOWING PRO 22 AND LEU 25,               1CBN 72
REMARK        4 RESPECTIVELY.                                                         1CBN 73
REMARK        5                                                                       1CBN 74
REMARK        5 IN SHEET *S1*, STRAND 3 IS QUESTIONABLY A STRAND.                     1CBN 75
REMARK        6                                                                       1CBN 76
REMARK        6 IN SHEET *S2*, STRAND 1 HAS NO HYDROGEN BONDS. THE                    1CBN 77
REMARK        6 BACKBONE ATOMS ARE IN BETA CONFORMATION.                              1CBN 78
SEQRES      1  46 THR THR CYS CYS PRO SER ILE VAL ALA ARG SER ASN PHE                 1CBN 79
SEQRES      2  46 ASN VAL CYS ARG LEU PRO GLY THR PRO GLU ALA LEU CYS                 1CBN 80
SEQRES      3  46 ALA THR TYR THR GLY CYS ILE ILE ILE PRO GLY ALA THR                 1CBN 81
SEQRES      4  46 CYS PRO GLY ASP TYR ALA ASN                                         1CBN 82
HET    EOH     66     5     ETHANOL                                               1CBN 83
FORMUL        2     EOH     C2     H6     O1                                      1CBN 84
HELIX         1 H1     ILE     7 PRO     19    1 3/10 CONFORMATION RESID 17-19     1CBN 85
HELIX         2 H2     GLU     23 THR     30    1 ALPHA-N DISTORTION AT START      1CBN 86
SHEET         1 S1     3 CYS     32 ILE     35     0                              1CBN 87
SHEET         2 S1     3 THR     1 CYS     4-1                                    1CBN 88
SHEET         3 S1     3 ASN     46 ASN   46-1                                    1CBN 89
SHEET         1 S2     1 THR     39 PRO     41   0                                1CBN 90
TURN          1 T1     ARG     17 GLY     20                                      1CBN 91
TURN          2 T2     PRO     41 TYR     44                                      1CBN 92
SSBOND          1 CYS     3 CYS     40                                            1CBN 93
SSBOND          2 CYS     4 CYS     32                                            1CBN 94
SSBOND          3 CYS     16 CYS     26                                           1CBN 95
CRYST1          40.763     18.492     22.333     90.00     90.61     90.00     P 21 2     1CBN 96
ORIGX1            1.000000     0.000000     0.000000     0.00000                  1CBN 97
ORIGX2            0.000000     1.000000     0.000000     0.00000                  1CBN 98
ORIGX3            0.000000     0.000000     1.000000     0.00000                  1CBN 99
SCALE1            0.024532     0.000000     0.000261     0.00000                  1CBN 100
SCALE2            0.000000     0.054077     0.000000     0.00000                  1CBN 101
SCALE3            0.000000     0.000000     0.044779     0.00000                  1CBN 102
ATOM          1 N ATHR       1     16.864     14.059     3.442     0.80     6.22     1CBN 103
ATOM          2 N BTHR       1     17.633     14.126     4.146     0.20     8.40     1CBN 104
ATOM          3 CA ATHR      1     16.868     12.814     4.233     0.80     4.45     1CBN 105
ATOM          4 CA BTHR      1     17.282     12.671     4.355     0.20     7.82     1CBN 106
ATOM          5 C THR        1     15.583     12.775     4.990     1.00     4.39     1CBN 107
ATOM          6 O THR        1     15.112     13.824     5.431     1.00     7.04     1CBN 108
ATOM          7 CB ATHR      1     18.060     12.807     5.200     0.80     5.42     1CBN 109
ATOM          8 CB BTHR      1     18.202     11.709     5.108     0.20     11.07    1CBN 110
ATOM          9 OG1ATHR      1     19.233     12.892     4.380     0.80     7.87     1CBN 111
...
ATOM          750     2HB      ASN     46     11.960     3.924     12.790     1.00     4.74     1CBN 852
ATOM          751     1HD2     ASN     46     12.053     4.334     16.647     1.00     7.82     1CBN 853
ATOM          752     2HD2     ASN     46     11.317     5.324     15.663     1.00     7.54     1CBN 854
TER           753     ASN     46                                                  1CBN 855
HETATM        754     C1     AEOH     66     14.823     −0.159     13.271     0.70     13.49     1CBN 856
HETATM        755     C1     BEOH     66     15.763     −0.521     12.803     0.30     10.99     1CBN 857
HETATM        756     C2     EOH     66     15.702     0.904     12.771     1.00     17.90     1CBN 858
HETATM        757     O     AEOH     66     15.029     2.134     12.501     0.70     7.21     1CBN 859
HETATM        758     O     BEOH     66     14.540     1.708     12.931     0.30     6.09     1CBN 860
CONECT        44     43     665                                                   1CBN 861
CONECT        54     53     546                                                   1CBN 862
CONECT        269     268     457                                                 1CBN 863
CONECT        457     269     456                                                 1CBN 864
CONECT        546     54     545                                                  1CBN 865
CONECT        665     44     664                                                  1CBN 866
CONECT        754     756                                                         1CBN 867
CONECT        755     756                                                         1CBN 868
CONECT        756     755     758                                                 1CBN 869
CONECT        757     756                                                         1CBN 870
CONECT        758     756                                                         1CBN 871
MASTER        66     0     1     2     4     2     0     6     757     1     11     4     1CBN 872
END                                                                               1CBN 873
```

FIG. 4. (*Continued*)

Acknowledgment

The author wishes to acknowledge the Judge Institute of Management Studies at the University of Cambridge, where this work was performed.

References

Boisot, M., Canals, A., & MacMillan, I.C.(2004). Simulating I-Space (SIS): An agent-based approach to modeling knowledge flows. Wharton Entrepreneurial Programs Working Paper Series. Retrieved August 2004, from http://www.wep.wharton.upenn.edu/Research/SimISpace3_200405.pdf

Boisot, M.H. (1995). Information space: A framework for learning in organizations, institutions and culture. London: Routledge.

Boisot, M.H. (1999). Knowledge assets: Securing competitive advantage in the information economy. Oxford: Oxford University Press.

Boisot, M.H., & Child, J. (1996). From fiefs to clans and network capitalism: Explaining China's emergent economic order. Administrative Science Quarterly, 41, 600–628.

Brinklow, T. (2004). Domains, ontologies, models and the knowledge creation cycle. Brighton Business School Occasional/Working Paper Series (BBSW04–3).

Doz, Y., Santos, J., & Williamson, P. (2001). From global to meta-national. Cambridge: Harvard Business School.

Polanyi, M. (1958). Personal knowledge. London: Routledge & Kegan Paul.

Politz, J., van Driel, R., Sauer, M., & Pombo, A. (2003). From linear genome to 3-d organization of the cell nucleus. Genome Biology, 4, 310.

DOI: 10.1002/asi