PROTEIN SIMILARITY SCORE: A SIMPLIFIED VERSION OF THE BLAST SCORE AS A SUPERIOR ALTERNATIVE TO PERCENT IDENTITY FOR CLAIMING GENUSES OF RELATED PROTEIN SEQUENCES

Christopher M. Holman†

TABLE OF CONTENTS

† 2004 Christopher M. Holman, Ph.D., 1993 (Biochemistry and Molecular Biology) University of California, Davis; J.D., 1998 Boalt Hall, University of California, Berkeley. Dr. Holman is Vice-President of Intellectual Property at PhyNexus, Inc. and chairs the AIPLA Biotechnology in the Courts Subcommittee. The views expressed herein are solely those of the author. The article does not necessarily state the views of PhyNexus, Inc., the AIPLA, nor any client or former client. The author wishes to thank Dr. Claes Gustafsson of DNA 2.0, Inc. and Brian Stanton of the U.S. Patent and Trademark Office for useful discussion and comments.

I. INTRODUCTION

Proteins constitute an important class of biological molecules, exhibiting a wide range of useful functional attributes. Many proteins form the basis of valuable commercial products, including therapeutics, diagnostics, research reagents, genetically modified organisms and industrial enzymes. In view of their commercial relevance, it is no surprise that the developers of novel proteins often seek patent protection for these molecules and the polynucleotides encoding them.[1] In recent years, as more protein inventions have matured into major commercial products, these patents have increasingly become the subject of enforcement and litigation. Oftentimes, however, these actions result in the patent being found invalid and/or not to cover the allegedly infringing product.[2] In a recent article aimed primarily towards the biotechnology community, Dufresne and Duval surveyed a large number of granted patents claiming genetic sequences and found the typical approaches to claiming such sequences to be "heterogeneous and imprecise, which may lead to questions regarding their validity."[3]

Clearly, from the point of view of the patentee, it is critical to employ a claiming strategy that will result in valid claims and maximal breadth of coverage. At the same time, the public has an interest in limiting protein claims to a reasonable scope, commensurate with an invention's contribution to the field. Furthermore, a claim reciting a genus of proteins should be sufficiently definite to adequately apprise third-parties of the metes

1. To obtain a rough measure of the extent to which protein sequences are the subject of patent claims, on May 19, 2004, the United States Patent and Trademark Office (PTO) database was searched for claims containing the terms "protein" or "polypeptide" and "sequence." There were 17,250 patents issued since 1976 that contained these terms. *See* USPTO, *at* http://www.uspto.gov.

2. *See*, *e.g.*, Chiron Corp. v. Genentech, Inc., 363 F.3d 1247 (Fed. Cir. 2004) (finding antibody claim invalid for failure to satisfy written description requirement); Noelle v. Lederman, 355 F.3d 1343 (Fed. Cir. 2004) (finding antibody claim invalid for failure to satisfy written description requirement); Biogen, Inc. v. Berlex Labs, Inc., 318 F.3d 1132 (Fed. Cir. 2003) (finding patents relating to recombinant production of human interferon not infringed); Regents of the Univ. of Cal.v. Eli Lilly and Co., 119 F.3d 1559 (Fed. Cir. 1997) (finding claims to genes encoding insulin invalid for failure to satisfy written description requirement); Genentech, Inc. v. Wellcome Found Ltd., 29 F.3d 1555 (Fed. Cir. 1994) (finding modified version of human tPA having several sizable deletions did not infringe claims to human tPA under the doctrine of equivalents); Amgen, Inc. v. Chugai Pharm. Co., 927 F.2d 1200 (Fed. Cir. 1991) (finding claims to DNA sequence encoding erythropoietin analogs not enabled).

3. Guillaume Dufresne & Manuel Duval, *Genetic Sequences: How Are They Patented?*, 22 NATURE BIOTECHNOLOGY 231, 231–32 (2004).

and bounds of the claims. In this article, I propose an approach to claiming proteins that addresses these issues in a manner that is in many ways superior to currently prevalent claiming strategies. The approach involves claiming a genus of related proteins sequences as defined by a reference sequence and a "similarity score," and is analogous to the standard technique of claiming proteins in terms of a reference sequence and percent identity.

I will begin by discussing some specific characteristics of proteins that make broad claim scope critical to the patentee, while at the same time rendering it difficult to obtain protein claims that are both valid and sufficiently broad to preclude design around by trivial modification. In the past, a number of approaches to claiming broad genuses of protein sequence have been employed; I will review several of these approaches and discuss some technical and legal obstacles with regard to their use. Next, I will describe what I refer to as the "similarity score" approach to claiming proteins, pointing out some of its advantages relative to the alternative approaches. Finally, I will explain how freely available software, such as the widely used Basic Local Alignment Search Tool, commonly known as BLAST, can be used to implement this claiming strategy, and provide examples of the use of BLAST to draft and analyze claims, both with respect to a potentially infringing sequence or in evaluating a claim with respect to the prior art.[4]

II. THE CHALLENGE OF ACHIEVING EFFECTIVE PATENT PROTECTION FOR PROTEINS

A. Protein Analogs and the Complex Relationship between Protein Structure and Function

Structurally, a protein can be defined as a biopolymer composed of one or more chains of amino acids in a specific order, *i.e.*, polypeptide chains.[5] In naturally occurring proteins, the order (or "sequence") of the amino acids is determined by the base sequence of nucleotides in the gene that codes for the protein.[6] For the purpose of

4. The novelty and nonobviousness of an invention are judged against everything publicly known before the invention, as shown in earlier patents and other published material. This body of public knowledge is called "prior art." See 35 U.S.C. § 102 (2004), which states, "A person shall be entitled to a patent unless . . . . " This language is followed by a series of definitions.

5. *See* THOMAS E. CREIGHTON, PROTEINS: STRUCTURES AND MOLECULAR PRINCIPLES Ch. 1, (W.H. Freeman & Co. 1983).

6. *See id.* Ch. 2.

this paper, I will focus on single chain proteins and make the simplifying assumption that all the amino acids that make up a protein are "standard" amino acids, *i.e.*, selected from the group of twenty amino acids encoded by the standard genetic code.[7] Proteins can be defined structurally in terms of their "sequence," that is, the sequence of the amino acids in the polypeptide chain.

In its native state, a protein typically folds to form a complex three-dimensional structure, with various functional groups of the amino acids positioned in a manner that enable the protein to perform its function.[8] It is the protein's sequence that dictates the three-dimensional structure, and hence the functionality of the protein.[9] However, there is some redundancy in this relationship between sequence and the structure/function of a protein.[10] As a consequence, oftentimes, proteins having similar but non-identical sequences form substantially equivalent three-dimensional structures and exhibit substantially identical function.[11] The more similar two sequences are, the more likely it is that they will be functional equivalents. However, in many cases substantial variations between sequences are tolerated without giving rise to any substantial difference in three-dimensional conformation or function.[12] In naturally-occurring proteins, for instance, sequence similarity between two proteins can be used to infer an evolutionary relationship, *i.e.*, that the proteins share a common ancestor, and that the sequences have diverged during the course of evolution.

As a corollary, typically, the sequence of a protein can be altered by substitutions at one or more amino acid positions without substantially affecting the protein's three-dimensional structure or function. Even relatively substantial sequence variations, such as substitutions at 50% or more of the amino acid positions, or the introduction of multiple deletions or insertions in the sequence, can at times be accommodated without substantially altering function.[13]

7. Naturally occurring proteins often include other amino acids, which can be encoded by alternate genetic codes or the result of post-translational modifications. Protein claims typically do not address this issue. For the purpose of not overly complicating this article, I will disregard these non-standard amino acids, although generally the rationale behind the proposed claiming strategy would apply even if non-standard amino acids were taken into account.

8. *See* CREIGHTON, *supra* note 5.

9. *Id*.

10. *Id*.

11. *See id.* Ch. 6.

12. *Id*.

13. *Id.*; BIOINFORMATICS FOR GENETICISTS Ch. 14 (Michael R. Barnes & Ian C. Gray eds., John Wiley & Sons, Ltd. 2003).

Bearing in mind that a typical protein comprises hundreds of amino acids, that the standard genetic code allows for nineteen possible amino acid substitutions at each position, and that typically a variety of substitutions, insertions and/or deletions can be accommodated without substantially altering protein function, it is evident that a genus of related proteins sharing similar sequence and conserved function (referred to herein as "analogs," or "functional variants" of one another) can be astronomical in size. This has important ramifications for the patenting of novel and commercially relevant proteins. Unless protein claims encompass these analogs, it can be relatively straightforward for a competitor to design around a claim simply by generating and commercializing one of these analogs.[14] This can be accomplished, *e.g.*, by isolating a naturally occurring homolog of the protein, or by engineering a non-naturally occurring variant using any of a wide range of recombinant techniques.[15] For example, a straight-forward approach to designing around a patent claiming a specific protein sequence would be to make a number of point mutants of the protein and screen these mutants to identify one that retains the desired function. Alternatively, a more sophisticated approach, such as DNA shuffling or molecular directed evolution, could be used to generate a variant having a large number of substitutions while still retaining the desired function.

14. *See, e.g.*, Schering Corp. v. Amgen, Inc., 222 F.3d. 1347, 1350–52 (Fed. Cir. 2000). Amgen avoided Schering's patent covering interferon-α by commercializing a consensus interferon-α, which Amgen generated synthetically after reviewing the sequences of all known interferon-α subtypes. The consensus sequence contains, at each position, an amino acid present in one or more known interferon-α subtype, but does not duplicate the amino acid sequence of any single interferon-α subtype and thus, does not correspond to any naturally occurring interferon subtype. *Id.*; *see also*, Antony L. Ryan & Roger G. Brooks, *Innovation vs. Evasion: Clarifying Patent Rights in Second-Generation Genes and Proteins*, 17 BERKELEY TECH. L.J. 1265, 1276–78 (2002).

15. This was alluded to in *Enzo Biochem, Inc. v. Gen-Probe Inc.*, 323 F.3d 956 (Fed. Cir. 2002) (Enzo II), wherein an expert testified that "astronomical" numbers of mutated variations of the deposited sequence would fall within the scope of the claims, and that such broad claim scope is necessary to adequately protect Enzo's invention from copyists who could otherwise make a minor change to the sequence and thereby avoid infringement while still exploiting the benefits of Enzo's invention. *Id.* at 966. Techniques for generating protein variants include site-directed mutagenesis and directed evolution methodologies. S*ee*, *e.g.*, Richard Fox et al., *Optimizing the Search Algorithm for Protein Engineering by Directed Evolution*, 16 PROTEIN ENGINEERING 589 (2003); Claes Gustafsson et al., *Putting Engineering Back into Protein Engineering: Bioinformatic Approaches to Catalyst Design*, 14 BIOTECHNOLOGY 366 (2003); Jeremy Minshull & Willem PC Stemmer, *Protein Evolution by Molecular Breeding*, 3 CHEMICAL BIOLOGY 284 (1999); Phillip A. Patten et al., *Application of DNA Shuffling to Pharmaceuticals and Vaccines*, 8 BIOTECHNOLOGY 724 (1997).

In practice, there are a number of non-patent barriers to entry that might come into play in this scenario, *e.g.*, in the case of a regulated product, such as a human therapeutic, the alteration in sequence might raise new regulatory issues that have already been resolved for the original molecule.[16] Nevertheless, in many instances, the protection provided by a patent that only covers a single sequence would be of very little commercial value if these other barriers to entry could be overcome. This would particularly be the case for proteins that are not subject to the strict regulatory scrutiny of agencies such as the FDA, *e.g.*, a recombinant enzyme being commercialized for industrial purposes or for use as research reagents.[17] Since a user of an industrial enzyme is only interested in exploiting the catalytic function of the enzyme, it would generally be straightforward for a competitor to engineer and commercialize a functional variant that avoids a narrowly claimed patent without raising any attendant regulatory issues.

In principle, an inventor of a novel protein could attempt to identify specific sequences corresponding to functional variants of the protein, and explicitly recite these in a claim. However, as discussed above, the number of potential functional variants would be astronomical, so even a claim listing many thousands of sequences would cover only a tiny fraction. To avoid literal infringement of the claim, a competitor would need merely to identify and commercialize one of the many analogs that are not identified in the claim.

B. Application of the Doctrine of Equivalents to Protein Claims

Of course, to some extent the doctrine of equivalents is available to expand claim coverage to include analogs incorporating "insubstantial" changes.[18] There are a number of cases where the

16. Lester M. Crawford, D.V.M., Ph.D., Acting Commissioner of Food and Drugs, Department of Health and Human Services, The Law of Biologic Medicine, Statement Before the Committee on the Judiciary/U.S. Senate (June 23, 204), *at* http://www.hhs.gov/asl/testify/t040623.html (discussing the regulatory hurdles that must be traversed in order to secure FDA approval for follow-on versions of protein-based drugs).

17. For examples of such products, including proteases that are used as additives in laundry detergents, see http://www.genencor.com/wt/gcor/clean, (last visited Sept. 25, 2004), and DNA polymerases used in PCR, see http://www.appliedbiosystems.com/catalog, (last visited Sept. 25, 2004).

18. Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co., 535 US 722, 731–33 (2002) (stating that under the doctrine of equivalents, "[t]he scope of a patent is not limited to its literal terms but instead embraces all equivalents to the claims described").

doctrine of equivalents has been applied to protein claims.[19] Nevertheless, while the doctrine is clearly available in some cases to expand the scope of protein claims, it is hard to predict the extent of sequence variation that would be permitted before the difference between the claimed and accused sequences becomes so substantial as to avoid the doctrine. The doctrine should be available to prevent a competitor from avoiding a patent simply by introducing a single point mutation of negligible functional significance.[20] However, what about the case where an accused analog incorporates multiple changes, perhaps resulting in a relatively minor change in three-dimensional structure or some alteration in function?

Most proteins possess multiple functions, so a determination that a protein's function has been substantially altered would, to some extent, depend upon how the protein's function or functions are defined. For example, what if alterations to a protein's sequence change a commercially irrelevant function without altering what are perceived to be the commercially relevant functions of the protein, or the alterations merely change an ancillary property such as pH dependence or temperature stability? What if the alterations cause some change in the magnitude of a functional characteristic, but the change is commercially insubstantial? It is an open question as to what extent the doctrine of equivalents would be available to the patentee in situations such as these.

As illustrated by the *Festo* decisions and other recent Federal Circuit case law, the trend is clearly towards a severely restricted application of the doctrine of equivalents.[21] In light of this trend, and the overall uncertainty surrounding this area of the law, the prudent practitioner will, to the extent possible, seek broad literal coverage rather than rely upon the doctrine of equivalents to ensnare infringers.

C. Claims Defining Proteins Solely in Terms of Function are Generally not Valid

In the early days of biotechnology patenting, claims often defined proteins simply in terms of function. Sometimes this was all that was known about the protein at the time the application was filed, particularly in cases where the DNA encoding the protein had yet to

19. *See*, *e.g.*, Amgen, Inc. v. Hoechst Marion Roussel, Inc., 314 F.3d 1313 (Fed. Cir. 2003); Genentech, Inc. v. Wellcome Found Ltd., 29 F.3d 1555 (Fed. Cir. 1994).

20. *Genentech, Inc.*, 29 F.3d at 1566–67.

21. Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co., 234 F.3d 558 (Fed. Cir. 2000), *vacated by* 535 U.S. 722 (2002), *remanded to* 344 F.3d 1359 (Fed. Cir. 2003).

be isolated. However, functional definitions also result in broad literal coverage, which on its face encompasses functional variants of the claimed protein. If the only limitation is function, then in principle the claim should encompass any and all variants that retain the desired function. Unfortunately for the patentee, these claims have not fared well in the courts, and the overwhelming trend is clearly toward a requirement of some sort of structural definition of the protein, or at least a physical description of the protein that goes beyond mere functional characterization.[22]

For example, in *Amgen, Inc. v. Chugai Pharmaceutical Co.*[23] claims reciting a DNA sequence encoding a protein having an amino acid sequence "sufficiently duplicative" of erythropoietin (EPO) to possess EPO's biological property of causing an increased production of red blood cells were found invalid for failure to enable[24] the full scope of the claims.[25] The court cited the "manifold possibilities" for changes to the structure of EPO "with attendant uncertainty as to what utility will be possessed by these analogs," and the failure of Amgen to identify "structural requirements for producing compounds with EPO-like activity." [26]

In *Ex parte Maizel*,[27] a patent application disclosed the amino acid sequence of a human B-cell growth factor. Claims reciting a DNA vector encoding a protein having the disclosed amino acid sequence or a "biologically functional equivalent thereof" were found invalid for lack of enablement.[28] The Board of Patent Appeals

22. Antibodies, which are themselves proteins, apparently represent an exception to this rule. For example, in *Noelle v. Lederman*, 355 F.3d 1343, 1349 (Fed. Cir. 2004), the court stated that

> based on our past precedent, as long as an applicant has disclosed a "*fully characterized* antigen," either by its structure, formula, chemical name, or physical properties, or by depositing the protein in a public depository, the applicant can then claim an antibody by its binding affinity to that described antigen.

Id.

23. 927 F.3d 1200 (Fed. Cir. 1991).

24. The enablement requirement refers to the requirement of 35 U.S.C. §112, ¶ 1, that the specification provide a description of how to make and use the invention without "undue experimentation." *See also*, *In re* Wands, 858 F.2d 731, 737 (Fed. Cir. 1988).

25. In many of the cases discussed herein, the claims at issue pertain specifically to DNA molecules encoding the functionally defined proteins. However, the rationale behind the decisions should apply equally to claims directed to proteins *per se*. Indeed, the claims typically define the claimed DNA in terms of the encoded protein, and the determination of invalidity hinges upon issues relating to adequate disclosure of the protein.

26. *Amgen*, 927 F.2d at 1214.

27. 27 U.S.P.Q. 2d (BNA) 1662 (B. Pat. App. Interferences 1992).

28. *Id.* at 1665.

analogized the rejected claims to a single means claims, and opined that the "problem with the phrase 'biologically functional equivalent thereof' is that it covers any conceivable means, *i.e.*, cell or DNA, which achieves the stated biological result while the specification discloses, at most, only a specific DNA segment known to the inventor."[29]

In *Fiers v. Revel*,[30] an interference count which purported to cover all DNA molecules coding for beta interferon was found not to comply with the written description requirement,[31] for such an invention "requires a precise definition, such as by structure, formula, chemical name or physical properties," not a mere wish or plan for obtaining the claimed invention.[32] The court held that knowledge of the chemical nature of the DNA is a prerequisite to an adequate written description (and this requirement can be inferred to apply to the description of other biological molecules such as proteins).[33]

Particularly compelling support for the proposition that at least some definition of structure is needed for a valid protein genus claim can be found in *Regents of the University of California v. Eli Lilly and Co.*[34] In that decision, claims directed to genes encoding mammalian insulin were found to be inadequately supported by the description of rat insulin cDNA, and hence invalid for failure to satisfy the written description requirement.[35] The court distinguished the rejected claiming strategy from the accepted practice of describing a broad chemical genus by means of a generic formula.[36] The rejected claims described the genus in terms of its function, and did not define any structural features commonly possessed by members of the genus that distinguish them from others.[37]

The *Eli Lilly* decision has been controversial, and is viewed by some as a novel and ill-advised interpretation of the written description requirement.[38] Nevertheless, it is widely cited as

29. *Id*.

30. 984 F.2d 1164, 1171 (Fed. Cir. 1993).

31. 35 U.S.C. § 112, ¶ 1 requires that the specification shall contain a written description of the invention. *See*, *e.g.*, Vas-Cath, Inc. v. Mahurkar, 935 F.2d 1555, 1560 (Fed. Cir. 1991).

32. *Fiers*, 984 F.2d at 1171.

33. *Id*.

34. 119 F.3d 1559, 1566 (Fed. Cir. 1997).

35. *Id.* at 1568–69.

36. *Id*.

37. *Id*.

38. Enzo Biochem, Inc. v. Gen-Probe Inc., 323 F.3d 956, 976, 979–81 (Fed. Cir. 2002) (Rader, J., dissenting).

precedent, particularly in cases pertaining to biological molecules and other chemical entities.[39] Indeed, *Eli Lilly* caused the U.S. Patent and Trademark Office (PTO) to reevaluate its own application of the written description requirement. In response to the decision, they published "Guidelines for Examination of Patent Applications Under the 35 U.S.C. 112, ¶ 1 'Written Description' Requirement"[40] and a "Synopsis of Application of Written Description Guidelines" (referred to herein as the "Guidelines").[41] The Guidelines apply the written description requirement, as articulated in *Eli Lilly*, to a number of hypothetical claiming scenarios, many involving proteins and other biological inventions. Interestingly, while the Guidelines are the product of an administrative agency, they have been relied upon in a number of Federal Circuit decisions and have to some extent taken on the mantle of de facto judicial precedent.[42]

For example, in *Enzo II*, the Federal Circuit took judicial notice of the Guidelines, stating that the DNA invention at issue would be valid if it complied with the written description requirement *as interpreted by the Guidelines*, and directing the lower court on remand to determine if the claimed subject matter was adequately described in a manner *consistent with the PTO guidelines*.[43] The Guidelines were cited for the proposition that a biological molecule can be claimed in terms of function only when "coupled with a known or disclosed correlation between [that] function and [a] structure" that is sufficiently known or disclosed.[44]

More recently, in *Noelle v. Lederman*,[45] the Federal Circuit characterized an example in the Guidelines directed to a hypothetical antibody claim as precedent, and relied upon that example to find the claim at issue invalid for failure to comply with the written description requirement.[46] The example was deemed precedent based on it having been mentioned in the *Enzo II* decision. However, the biological molecules at issue in *Enzo II* were nucleotide sequences,

39. *See, e.g.*, *id.* at 964–65; Noelle v. Lederman, 355 F.3d 1343 (Fed. Cir. 2004); Univ. of Rochester v. G.D. Searle & Co., 358 F.3d 916 (Fed. Cir. 2004); Chiron Corp. v. Genentech, Inc., 363 F.3d 1247 (Fed. Cir. 2004).

40. Guidelines for Examination of Patent Applications Under the 35 U.S.C. §112, ¶ 1, "Written Description" Requirement, 66 Fed. Reg. 1099 (Jan. 5, 2001).

41. *See Synopsis of Application of Written Description Guidelines*, *at* http://www.uspto.gov/web/menu/written.pdf (last visited May 25, 2004).

42. *See Enzo*, 323 F.3d 956; *Noelle*, 355 F.3d 1343.

43. *Enzo*, 323 F.3d at 967.

44. *Id.* at 964 (emphasis omitted).

45. *Noelle*, 355 F.3d 1343.

46. *Id.* at 1349.

not antibodies or even proteins. In fact, *Enzo II* only referred to the antibody example in a single sentence, and merely as an example of the PTO's approach to written description analysis.[47] Holding that the antibody example from the Guidelines is "precedent" suggests that at this point the court might consider the entire Guidelines to constitute precedential authority.

The Guidelines themselves provide some guidance as to the degree to which specific sequence information must be provided in order to satisfy the written description requirement. For instance, in Example 11 of the Guidelines, a claim to "[a]n isolated allele of [a disclosed DNA sequence]" is found to be invalid for a number of reasons. [48] For one reason, there is no description of the actual sequence variations that might exist in nature. In addition, the genus would include members that would be expected to have widely divergent function, *i.e.*, the structure and function of one allele does not provide guidance to the structure and function of others.[49]

In Example 13 of the Guidelines, a claim to an "isolated variant of [a protein having a disclosed sequence]" is found to be invalid.[50] The rationale behind this finding includes the failure of the specification and claims to indicate distinguishing attributes shared by members of the genus, *i.e.*, the failure to identify structural features that could distinguish compounds in the genus from others in the protein class, and failure to place any limit on the number of amino acids substitutions, deletions, insertions and/or additions that could be made in a variant.[51]

However, in Example 14 of the Guidelines, a claim to variants of a disclosed protein was found to be valid when the claim was limited to variant sequences that are at least 95% identical to the disclosed sequence and are able to catalyze the reaction A → B (a functional attribute of the disclosed sequence).[52] The Guidelines' analysis of the example found that procedures for making variants which have 95% identity and which retain the functional activity are conventional in the art. It also found that there was no substantial variation amongst members of the genus since all the variants must possess the specified catalytic activity and must have at least 95% identity to the reference

47. *Enzo*, 323 F.3d at 964.

48. *See Synopsis of Application of Written Description Guidelines*, *supra* note 41, at 41.

49. *Id*.

50. *Id*. at 50.

51. *Id*.

52. *Id* at 41.

sequence.[53] There is no indication of how the 95% identity threshold was arrived at, nor as to how low the percent identity term could be varied without resulting in an overly broad and invalid claim.

To summarize the case law and the PTO's interpretation of that case law, Section 112 clearly requires more than a mere functional definition of a genus of proteins.[54] While it is possible to claim a genus of protein variants sharing similar sequences and common functionality, it is insufficient to merely recite a specific sequence and claim it and its functional variants. Rather, the degree of acceptable sequence variations should be specifically recited, so as to clearly delineate the metes and bounds of the claim in terms of structure, *e.g.*, the percent identity claim of Example 14 in the Guidelines.

Moreover, the sequence definition must to some extent be predictive of conserved function, though it need not be exactly predictive of function. For example, in Example 14, the 95% identity limitation would certainly encompass a large number of non-functional mutants of the disclosed protein. Nevertheless, 95% identity is predictive of function in the sense that it is much more likely that a sequence that is 95% identical to the disclosed protein will share a common function than would a randomly generated sequence. For that matter, it is more likely that a 95% identical sequence will share the function than a similar sequence having a lower percent identity, *e.g.*, 90% identity. In the Guidelines this principle is expressed in terms of it being "conventional in the art to find functionally active variants sharing 95% identity." Implicitly, there must be a point where the percent identity limitation becomes so low that it would no longer be "conventional" to find variants at that level of percent identity which retain the activity, and/or at which point the degree of sequence variation would constitute "substantial variation." At that point, the claim would presumably fail to comply with the written description requirement.

As a corollary, the more predictive the sequence definition is of conserved function, the more likely it is that the claim will satisfy the written description requirement.[55] Consider the 95% identity limitation from the Guidelines. For a protein sequence of typical length, the number of possible sequence variants that would share

53. *Id*.

54. 35 U.S.C. § 112, ¶ 1 (2004).

55. *See* David E. Huizenga, Comment, *Protein Variants: A Study on the Differing Standards for Biotechnology Patents in the United States and Europe*, 13 EMORY INT'L L. REV. 629, 655–66, (1999) (noting that predictability was a central issue in *Amgen v. Chugai*, and that "[p]rotein variant claims are particularly susceptible to the 'predictability' sword")

95% identity is enormous. Of this genus of molecules, a significant percentage would retain the function, *i.e.*, the limitation is to some extent predictive of conserved function. However, consider the same claim except that the percent identity limitation has been decreased to 20% identity. The number of potential variants sharing 20% identity would be much larger than the 95% identity group, because there are vastly more possibilities for varying the sequence. However, only a vanishingly small number of these variants would be expected to retain the function. It would likely require screening an inordinate number of these variants to find a functional analog, *i.e.*, identifying such a variant would no longer be "conventional in the art." Under the Guidelines, this large variation in sequence would likely be considered substantial, and the claim not to be in compliance with the written description requirement (nor the enablement requirement, to the extent that the amount of screening required constitutes undue experimentation).[56]

Let us assume that for a particular protein, the 95% identity limitation will satisfy the written description requirement while the 20% identity limitation will not. We must conclude that there is some threshold between 20% and 95% identity that must be exceeded in order for the claim to be valid. We have little guidance as to the exact magnitude of the threshold, but clearly it is based on the extent to which the percent identity limitation is predictive of homology and conserved function. This imposes a fundamental limitation on the breadth of claims of this type, since the size of the claimed genus is inversely proportional to the magnitude of the percent identity limitation. Clearly, the objective should be to employ a sequence limitation that covers the broadest range of functional sequence variants while including a minimal number of non-functional variants. This objective is best realized by using a sequence limitation that is maximally predictive of conserved function. The similarity score approach presented in this article approaches that ideal.

D. Defining a Protein Genus in Terms of Percent Identity

Before delving into the similarity score approach, let us review the more conventional percent identity approach to claiming a genus of proteins. Many thousand of examples of such claims appear in issued U.S. patents.[57] For example, see claim 15 of U.S. Patent No. 6,657,047 (the "'047 patent"):

56. 35 U.S.C. § 112 (2004).

57. *See* USPTO patent database, *at* http://www.uspto.gov.

> An isolated protein comprising an amino acid sequence 80% or more identical to a polypeptide encoded by amino acid residues 1 to 385 of SEQ ID NO:2.[58]

Claims in this format generally recite a "reference sequence" (*e.g.*, amino acid residues 1 to 385 of SEQ ID NO:2) and a specified percent identity (*e.g.*, 80% or more identical), thereby identifying a genus of polypeptides sharing some minimal threshold of sequence identity with one another. Most patent applications containing percent identity claims will include in the specification some definition of the term "identical." A typical definition, such as that provided in the '047 patent, will specify that the percent identity between a reference sequence and a query sequence (*i.e.*, a sequence being analyzed to determine whether it falls within the scope of the claim) is determined by aligning the sequences so that the highest order match is obtained, and comparing the aligned amino acids. The number of exact matches as a percent of the total number of amino acids in the reference sequence is determined, and this is the percent identity of the two sequences.[59] The determination is illustrated below in Example 1, where two short ten-amino-acid peptides are aligned.[60] Note that the sequences differ only at the fourth and sixth amino acid positions. Using the percent identity approach, we would say that since eight out of the ten positions are identical the peptides are 80% identical.

Example 1

```
M G E T Y F P L S A
| | |   |   | | | |
M G E S Y T P L S A
```

Note that accurate scoring is dependent upon the proper alignment of the sequences. Though often not explicitly stated, the logical definition of the appropriate alignment, including positioning of gaps, should be the alignment that results in the highest percent identity between the sequences. For very similar sequences, alignment is trivial. However, for more distantly related sequences, particularly when there are deletions in one or both of the sequences, the alignment is not so straightforward. Patent specifications

58. U.S. Patent No. 6,657,047 (issued Dec. 2, 2003).

59. *See* U.S. Patent No. 6,657,047, col. 3, l. 60 to col. 4, l. 63 (issued Dec. 2, 2003).

60. Throughout this article the standard single-letter symbols for amino acids are used.

typically include references to algorithms and/or computer programs for performing the alignment.[61]

To my knowledge, the validity of a percent identity claim has not been addressed in a reported judicial decision. However, as discussed above, the approach has been sanctioned by the PTO in the Guidelines, and the Federal Circuit has shown a marked deference to these Guidelines. Therefore, it is not unlikely that the Federal Circuit would look to the Guidelines when assessing the validity of a percent identity claim.

Note that the claim from the '047 patent does not include a functional limitation. Protein genus claims lacking a functional limitation will almost certainly encompass a large number of non-functional variants, because there are invariably amino acid positions that cannot be altered without disrupting function.[62] When amino acids are altered at multiple positions in a sequence the likelihood of an impact on function increases, *e.g.*, changing up to 20% of the residues in a protein (as permitted by an 80% identity claim) would in most cases result in an impaired-function mutant. In general, the lower the magnitude of the percent identity limitation (*i.e.*, the broader the claim) the higher percentage of non-functional variants predicted to fall within the bounds of the claim. This in turn raises the issue of utility, since a polypeptide lacking any known functional activity would likely fail to satisfy the utility requirement.[63] While the inadvertent recitation of some non-functional species does not necessarily invalidate a genus claim, there could be a point where an excessive number of non-functional species raises validity issues, particularly where the specification supplies no structural guidance to distinguish functional from non-functional species.[64]

61. U.S. Patent No. 6,657,047 col. 3, l. 60 to col. 4, l. 24 (issued Dec. 2, 2003).

62. A classic example would be the family of proteases referred to as the serine proteases (*e.g.*, subtilisin), wherein it is known that any mutation that disrupts the amino acids that make up the enzyme's "catalytic triad" will severely disrupt the catalytic ability of the enzyme.

63. The utility requirement arises from 35 U.S.C. § 101, which has been interpreted as requiring that in order to be patentable an invention must have a substantial practical utility. *See, e.g.*, Brenner v. Manson, 383 U.S. 519 (1966); *In re* Ziegler, 992 F.2d 1197 (Fed. Cir. 1993). A deficiency in utility under 35 U.S.C. § 101 also creates a deficiency under 35 U.S.C. § 112, ¶ 1. *See In re* Brana, 51 F.3d 1560 (Fed. Cir. 1995).

64. *See, e.g.*, Amgen, Inc. v. Chugai Pharm. Co., 927 F.3d 1200, 1214 (Fed. Cir. 1991) (finding claims non-enabled in view of the "manifold possibilities" for changes to the structure of EPO "with attendant uncertainty as to what utility will be possessed by these analogs," and the failure of Amgen to identify "structural requirements for producing compounds with EPO-like activity").

In most percent identity claims the issue of non-functional species is addressed by means of a functional limitation, which explicitly limits the claim to functional variants. Example 14 of the Guidelines, or claim 1 of U.S. Patent No. 6,667,391 states:

> An isolated polypeptide comprising an amino acid sequence that is a least 99% identical to SEQ ID NO: 23, wherein the polypeptide exhibits stem cell growth factor activity.[65]

This claim format is the currently preferred approach to claiming a genus of structurally related proteins.[66] The percent identity limitation provides a definite structural recitation of the claimed molecules, and the functional limitation explicitly excludes molecules lacking utility from the claimed genus. For the remainder of this article, the term "percent identity claim" will refer to a claim that includes percent identity and functional limitations.

As discussed above, the rationale underlying the percent identity approach is that the percent identity between two sequences is predictive of conserved function and sequence homology.[67] But what exactly is the relationship between percent identity, conservation of function, and sequence homology? The term "percent homology" is often used, but this terminology tends to confuse two distinct issues. Homology is a biological term denoting that two protein sequences have evolved from a common ancestor. Depending on the evolutionary distance between the sequences, the number of amino acid differences between homologous proteins can be great, at some point rendering it difficult or impossible to discern the common ancestry of the sequences.

Thus, strictly speaking, the term "percent homology" is a misnomer; either two proteins can trace their ancestry back to a common sequence, or they cannot. However, percent *identity* is a predictor of homology; in general, the extent of percent identity between two sequences roughly correlates with the probability that the sequences are homologous. This is premised on that fact that, in view of the vast number of possible amino acid sequences, it is extremely unlikely that two different protein sequences sharing substantial percent identity would have arisen independently during the course of evolution, and hence, they must in all likelihood be

65. U.S. Patent No. 6,667,391 (issued Dec. 23, 2003).

66. See USPTO patent database, at http://www.uspto.gov, for examples of recent patent and published patent applications.

67. *See Synopsis of Application of Written Description Guidelines*, *supra* note 41, at 46.

descended from the same common ancestor.[68] As a corollary, it is likely that homologous proteins will share function, because to the extent the function confers some advantage upon an organism, there will be evolutionary pressure to maintain that function, *i.e.*, loss-of-function mutations tend to be selected against.[69] Hence, we have the relationship between homology, percent identity and conserved function in naturally occurring proteins: proteins sharing high percent identity are likely homologous, and they share conserved function because otherwise they would not have been retained by an organism during the course of evolution.

Note that this relationship breaks down to some extent in the case of non-naturally occurring proteins, *i.e.*, synthetic sequences generated using recombinant technology. For example, the genus of all possible protein sequences sharing 80% identity with a given reference sequence is astronomical, and it is likely that only a small fraction of these actually exist in nature as homologs of the reference sequence.[70] However, in principle any of the sequences falling within the genus could be synthesized using recombinant technology. Hence, unless the claimed genus is limited to naturally occurring proteins, it will encompass a vast number of sequences that are technically not homologs. Patent claims typically lack any such limitation, although in some cases it might be implicit, *e.g.*, in the case of a percent identity claim that lacks a functional limitation.

Furthermore, a percent identity delimited genus including synthetic variants will generally include a large number of non-functional proteins, because without the constraint of evolutionary pressure there is no mechanism to select against loss of function.[71] At the same time, the genus will include many synthetic variants that

68. *See* CREIGHTON, *supra* note 5, Ch. 12.

69. STRUCTURAL BIOINFORMATICS Ch. 12, (Philip E. Bourne & Helge Weissig eds., Wiley-Lis, Inc. 2003).

70. For example, consider a typical protein 300 amino acids in length. Assuming 20 possible amino acid residues, there are 19 possible changes at each of the 300 positions. The equation to calculate the number of possible 300 amino acid sequences sharing 80% identity is (19 x 300) + (19 x 300)(19 x 299) + (19 x 300)(19 x 299)(19 x 298) + . . . , continuing up to the point where the equation includes the term containing (19 x 240), *i.e.*, 240 is 80% of 300). The solution to this equation is 1.8×10^{226}, *i.e.*, there are 1.8×10^{226} possible variants sharing 80% identity with any given 300 amino acid protein. On the other hand, the number of different species of living organisms has been estimated at from 10^6 to 10^8, a miniscule fraction of the total number of possible 80% identical variants. *See* James Cotton, *Re: How Many Different Living Organisms Are There Today?*, MadSci Network, *at* http://www.madsci.org/posts/archives/jun2000/959840635.Zo.r.html (last visited Sept. 25, 2004).

71. STRUCTURAL BIOINFORMATICS, *supra* note 69.

retain function but that are not "homologs" in the biological sense since they are not literally descended from a common ancestor. As discussed above, typically there are many ways in which a sequence can be synthetically altered without disrupting activity; indeed, in many cases the function can actually be improved.[72] As a consequence, percent identity is much less effective at predicting conserved function when the claim is not limited to naturally occurring proteins. The generation of synthetic proteins, often times with novel and/or improved function, is a very active field of endeavor, with a number of companies seeking to commercialize such proteins.[73] Furthermore, the generation of synthetic analogs would be an obvious tactic for avoiding claims directed to a product based on a naturally occurring protein sequence. Clearly, the prudent practitioner will do well to draft protein genus claims with these potential synthetic variants in mind.

III. THE SIMILARITY SCORE APPROACH TO CLAIMING PROTEIN GENUSES

At this point, I will present and explain the similarity score approach to claiming a protein genus. Compared to percent identity claims, the similarity score approach has the following advantages: (1) it better accounts for the fact that not all amino acid substitutions are functionally equivalent, some being more conservative than others; (2) it better accounts for insertions and deletions, which are typically treated no differently than substitutions in percent identity approaches; (3) it takes into account the length and structural complexity of a sequence; (4) it is particularly well suited for use with synthetic, non-naturally occurring sequences; and (5) it is more consistent with the manner in which scientists evaluate related sequences for homology and/or conserved function. Furthermore, the scope of similarity score claims is at least as definite and unambiguous as a percent identity claim. The determination of whether a sequence of interest falls within the claimed genus can be accomplished in a straightforward manner using simple arithmetic, with or without the aid of a computer. The software needed to automate the determination is freely available over the Internet, *e.g.*,

72. *See supra* note 15.

73. Examples include Maxygen, Inc. (Redwood City, CA), Applied Molecular Evolution, Inc. (San Diego, CA) and Diversa Corporation (San Diego, CA).

the BLAST algorithms at the National Center for Biotechnology Information (NCBI) website.[74]

A. Comparing Protein Sequences in Terms of Similarity Score

To begin, I will demonstrate how a similarity score is determined for a pair of aligned sequences. I will then work through a number of examples, pointing out the advantages of the approach compared to percent identity. Finally, I will demonstrate how the BLAST alignment tool can be used to draft and analyze similarity score claims.

Similarity scores for aligned sequences are widely used in a number of computer-implemented approaches to protein sequence analysis, including the widely used BLAST sequence alignment tool.[75] Basically, a score for two aligned sequences is determined by means of a twenty-by-twenty scoring matrix, representing the 210 possible pairings of the twenty amino acids encoded by the standard genetic code. A number of different scoring matrices have been derived, many having attributes that make them particularly well suited for analyzing particular types of sequences and alignments.[76] In the following examples, I will employ exclusively the BLOSUM62 matrix (see Figure 1 below), a good general-purpose scoring matrix and the default used in the NCBI version of BLAST.[77] The pairings represent amino acids that line up with one another in a given sequence alignment. The score for any pair can be positive or negative, with identical amino acid pairs (representing a position in the alignment that is conserved between the two sequences) having the highest scores, followed by those that share some degree of homology (*e.g.*, leucine and isoleucine), with the more non-conservative pairings having the most negative scores. The more positive the score, the more similar the sequences and the more likely it is that they are homologous and/or share a conserved function.[78]

74. A general overview of scoring matrices and the BLAST sequence alignment tools are provided at the National Center for Biotechnology Information (NCBI) BLAST website, *at* http://www.ncbi.nlm.nih.gov/BLAST (last visited May 25, 2004). The BLOSUM62 matrix is described in Steven Henikoff & Jorja G. Henikoff, *Amino Acid Substitution Matrices from Protein Blocks*, 89 PROC. NAT'L ACAD. SCI. 10915 (1992), *available at* http://www.pnas.org/cgi/reprint/89/22/10915.pdf (last visited Oct. 5, 2004).

75. NCBI BLAST, *supra* note 74.

76. *Id*.

77. *Id*.

78. *Id*.

Consider the case of two short sequences, where there is no gap in either of the aligned sequences throughout the length of the sequences being compared. Example 2 is a fifteen amino acid sequence (the N-terminal of maltoporin precursor, Accession No. NP_807741.1) aligned with itself, *i.e.*, an alignment of 100% identical sequences.

Example 2

```
M M I T L R K L P L A V A V A
M M I T L R K L P L A V A V A
5 5 4 5 4 5 5 4 7 4 4 4 4 4 4  (68)
```

Underneath each amino acid is the score for that particular pairing, obtained by finding the number at the intersection of the amino acids on the BLOSUM62 matrix (Fig. 1). The score for the alignment is sixty-eight, the sum of the fifteen individual scores. Note that a score of sixty-eight is the highest possible score that any amino acid sequence could generate when aligned with this particular fifteen amino acid segment, because the score for any non-identical pair is always lower than the score for an identical pairing.

Example 3 is another fifteen amino acid sequence taken from maltoporin precursor, again aligned with itself.

Example 3

```
R F Y Q R H D V H M I D F Y Y
R F Y Q R H D V H M I D F Y Y
5 6 7 5 5 8 6 4 8 5 4 6 6 7 7 (89)
```

Figure 1

BLOSUM62 Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Note that this alignment generates a score of eighty-nine (twenty-one points higher than the score from Example 2). This reflects the facts that this sequence includes residues that generate higher score for identical pairings, *e.g.*, H, Y, F and D. Residues that are the most chemically and/or structurally unique, and/or that appear less frequently in proteins, such as W, C, H, P and Y produce the highest scores for identical matches (eleven, nine, eight, seven and seven, respectively), because the fact that they are conserved between two sequences is a more significant indication of homology than the conservation of more commonly occurring and/or more easily substituted for amino acids, such as A, I, L, S and V (all generating score of four). Note that the theoretical maximum score for a fifteen amino acid alignment is 165 (for a sequence of fifteen consecutive Ws), and the theoretical minimum score is sixty (for a sequence consisting only of A, I, L, S and/or V).[79]

Using a percent identity approach, both of these alignments would be scored as 100% identical, the highest possible score, even though the identity of the second pair is substantially more indicative of homology than it is for the first pair. In fact, the significance of 100% identity to any given sequence depends upon the amino acid composition of the sequence. The identity is much more predictive of homology when the sequence has a large percentage of difficult to substitute for amino acids (such as W and C) compared to a sequence that is rich in easily substituted for amino acids (I, L , etc.).

Example 4 depicts the fifteen amino acid sequence of Example 3 aligned with a similar but non-identical sequence.

Example 4

```
R F Y Q R H D V H M I D F Y Y
R Y Y Q R H D L H I I D Y F Y
5 3 7 5 5 8 6 1 8 1 4 6 3 3 7 (72) 10/15 67% identity
```

Only ten of the fifteen pairings match, so the sequences are 67% identical. In calculating the similarity score, the ten identical residues are scored the same as in Example 3, while the scores for the other five pairs are less, *e.g.*, the F-Y pair generates a score of three, while the F-F pair generates a score of six; the V-L pair generates a score of one, while V-V generates a score of four. Note that in this example,

79. *Id.*; BIOINFORMATICS FOR GENETICISTS, *supra* note 13, Ch. 12.

all of the substitutions are relatively conservative, with the substituted amino acids sharing similar chemical and structural characteristics. As a result, all of the scores are positive, albeit lower than the corresponding identical pairing scores. The resulting score for the alignment is seventy-two, necessarily less than the theoretical maximum score of eighty-nine generated in Example 3, but still higher than the score of sixty-eight generated for the 100% identical sequences in Example 2.

Example 5 is another alignment of the sequence of Example 3 with a different 67% identical sequence.

Example 5

```
R F Y Q R H D V H M I D F Y Y
R Y Y Q R H D L H I I D T K Y
5 3 7 5 5 8 6 1 8 1 4 6-2-2 7 (62) 10/15 67% identity
```

In this case, two of the substitutions are non-conservative, *i.e.*, F-T and Y-K. Each of these substitutions generates a negative score (–2). As a result, the similarity score for this alignment is only sixty-two , substantially lower than the score of seventy-two generated in Example 4 for another 67% identical alignment. This illustrates the imprecision of percent identity as a predictor of homology. Because percent identity does not take into account the chemical nature of the substituted amino acids and the extent to which a change is conservative, two alignments of the same length and percent identity can generate very different similarity scores depending upon the nature of the sequences and the substitutions. These differences in score represent differences in the likelihood that the sequences are homologous and share common functional/structural characteristics. In this regard then, similarity score is a much better measure of homology than percent identity, and as such provides better linkage between structure and function for purposes of drafting valid patent claims.

Example 6 is identical to the alignments of Examples 4 and 5, except that lower sequence has a deletion at the positions corresponding to amino acids thirteen and fourteen in the upper sequence.

Example 6

```
R F Y Q R H D V H M I D F Y Y
R Y Y Q R H D L H I I D - - Y
5 3 7 5 5 8 6 1 8 1 4 6(-13)7 (53)
```

In calculating a similarity score, deletions, *i.e.*, gaps, are scored using the formula y=a+bx, where y is the score for the gap, a is a gap existence penalty, b is a gap extension penalty, and x is the length of the gap (in this example the gap has a length of two, corresponding to the two unpaired residues in the query sequence). The magnitude of the gap penalties can be varied, in the same way that different scoring matrices can be employed. In this article, a gap existence penalty of –11 and gap extension penalty of –1 will be used. These are typical values, and are the default penalties used in NCBI BLAST. In Example 6, the two amino acid gap generates a score of –13. The similarity score for the alignment is fifty-three, significantly lower than for the 67% identity alignments that did not include any gap.

The lower score for the alignment of Example 6 compared to Examples 4 and 5 reflects the biological significance of the insertion of a gap in the alignment. From an evolutionary and functional conservation standpoint, a gap in an alignment generally represents a much more significant difference between sequences than a corresponding amino acid substitution. The presence of a gap, no matter how small, should be weighted much more heavily than a simple substitution when analyzing aligned sequences for homology. At the same time, once a gap has been introduced, the incremental extension of the gap is only slightly more predictive of lack of homology. In other words, the introduction of a short one or two residue gap should result in a relatively large negative hit to the similarity score.[80] But most of the disruption is simply a consequence of the insertion of a gap; the difference between a gap of two residues and a gap of four residues is only minimal. The gap existence and gap extension penalties normally employed account for this phenomenon by imposing a large penalty (in this case –11) for the introduction of the gap, but only a small penalty (in this case –1 per residue) for the incremental extension of the gap.

Compare this treatment of deletions with the percent identity approach. Most patent specifications that define "percent identity" treat a deletion the same as any other mismatch. Thus, a one amino

80. *See* NCBI BLAST, *supra* note 74.

acid deletion is penalized to the same extent as a conservative amino acid substitution, and a ten amino acid deletion is penalized ten times as much as a one amino acid deletion. This is clearly a gross oversimplification, since a one amino acid deletion is predicted to be much more disruptive to function than a conservative substitution, but extension of the deletion to ten amino acids only marginally increases the damaging effect of the initial introduction of the gap.

In summary, Examples 4–6 provide three alignments, all of which would generate identical scores under a percent identity approach. The similarity scores vary from seventy-two to sixty-three to fifty-three, depending upon the nature of the non-identity. These different scores reflect the fact that the sequences in Example 4 are substantially more likely to be homologous than are those in Example 5, which in turn are more likely to be homologous than the sequences of Example 6. If the intent in claiming a genus of related protein sequences is that the sequences be homologous and/or share structural and functional characteristics, the similarity score approach is clearly superior to the percent identity approach.

Another advantage of the similarity score approach is that it explicitly accounts for the lengths of the aligned sequences. A high degree of conservation over a long stretch of sequence is more predictive of homology than is the same degree of conservation over a short segment. For example, it is intuitively obvious that 80% identity between two 500 amino acid sequences is much more predictive of homology than 80% identity of two, ten amino acid long segments. However, while intuitively obvious, a simple percent identity approach fails to account for this distinction. The similarity score approach explicitly accounts for sequence length, rendering a proportionately higher score to longer regions of conserved sequence.[81] To illustrate, refer to Example 7, where the length of the alignment in Example 4 is doubled simply by joining the aligned sequences in tandem to a copy of itself. The percent identity remains 67%. The similarity score, on the other hand, doubles to 144, reflecting the fact that it is much more likely that two thirty-amino-acid sequences sharing this degree of similarity are homologous than would be the case for the corresponding fifteen-amino-acid sequences.

81. BIOINFORMATICS FOR GENETICISTS, *supra* note 13, Ch. 4.

Example 7

```
RFYQRHDVHMIDFYYRFYQRHDVHMIDFYY
RYYQRHDLHIIDYFYRYYQRHDLHIIDYFY
5375586181463375375586181463 37 (144)20/30 67% identity
```

Because the similarity score approach more accurately accounts for the functional impact of specific amino acid substitutions and gaps in an alignment, it is particularly suited for predicting functional synthetic analogs. As a consequence, a genus of related sequences as defined by similarity score will include a substantially higher percentage of functional variants compared to genus of comparable size defined by percent identity. Assuming that some threshold predictive accuracy is required of a sequence limitation, *i.e.*, some minimal fraction of a defined genus must be functional in order for the claim to be valid, it follows that the similarity score approach enables valid claims encompassing a substantially higher number of functional analogs than could be achieved by percent identity. For the patentee, this translates into expanded literal claim coverage.

Another advantage of the similarity score approach is that it is more consistent with the approach used by scientists to compare sequences. Typically, a scientist evaluating a sequence and looking for homologs or functionally-related molecules will use a sequence alignment tool such as BLAST. The output of a BLAST search is a list of related sequences, ranked in order of similarity and including scores that represent the likelihood that the sequences are related to the query sequence.[82] These scores are all derived from similarity scores based on a substitution matrix and gap penalty.[83] A similarity score is a technically superior approach to sequence comparison, and will be viewed by a biologist as a more rational approach to claiming a protein genus than the percent identity approach.

B. Protein Claims Reciting Similarity Score

To employ similarity scores in the claiming of proteins, I propose simply using a modified version of the percent identity claim

82. This is the standard alignment tool provided by the NCBI, which is the primary public resource for molecular biology information in the U.S. *See* NCBI, *at* http://www.ncbi.nlm.nih.gov (last visited Oct. 4, 2004).

83. For a more detailed explanation of BLAST scoring, see the documentation provided at the NCBI BLAST website, *at* http://www.ncbi.nlm.nih.gov/BLAST (last visited May 25, 2004).

format, substituting similarity score for the percent identity limitation. For example:

> An isolated polypeptide comprising an amino acid sequence that when optimally aligned with SEQ ID NO:1 will generate a similarity score of at least X using the BLOSUM62 matrix, a gap existence penalty of 11, and a gap extension penalty of 1, wherein the polypeptide has Y functional activity.

The claim has a functional limitation to exclude non-functional analogs and a similarity score limitation precisely delimiting the scope of the claim. It also specifically recites the scoring matrix and gap penalties; when using a similarity score approach to claiming it is critical that these terms be explicitly defined, either in the claims themselves or in the definition of similarity score provided in the written description. The scoring matrix and gap penalties used in this example are the NCBI BLAST defaults, and are probably the best defaults to use in general claim drafting.[84]

The generation of an unambiguous score for two sequences depends upon the optimal alignment being unambiguous. The term "optimal alignment" should be defined simply as the alignment (including the introduction of gaps in the sequences as necessary) that results in the highest similarity score. Some issued patents provide complex definitions of alignment that require the use of a computer-implemented algorithm to determine optimal alignment.[85] This is not necessary with the similarity score approach. Practically speaking, a computer algorithm might be required in some cases to initially figure out what the optimal alignment is (when the sequences are quite divergent and/or when gaps must be introduced), but a computer is not required to define optimal alignment or to determine whether an accused sequence falls within the scope of the claim.

For example, during litigation a patentee would simply present to the court an alignment of the claim's reference sequence and the accused sequence (which would likely have been generated using a computer) and tabulate the similarity score (which can be done

84. In some cases, a more sophisticated approach might involve using alternative scoring matrices and/or gap penalties that are more biologically relevant for the particular genus of claimed sequences. In principle, this would provide some (likely marginal) improvement in the accuracy of the similarity score in predicting functional variants. However, most practitioners would probably simply employ the suggested default parameters, which are on average the best for typical sequences; any predictive improvement achieved by optimizing the scoring system would likely be minimal compared to the substantial improvement that results simply in going from percent identity to a similarity score approach.

85. *See e.g.*, U.S. Patent No. 6,605,450, claim 1 (issued Aug. 12, 2003).

manually as described above). If the score exceeds the score recited in the claims, the sequence literally infringes (assuming that any functional or other limitations are also satisfied). The result is unambiguous, and the accused infringer will have no basis for arguing for a different alignment because by definition the correct alignment is the one that yields the highest score. Any alternative alignment proposed by the accused infringer would either generate a lower score, and hence by definition would not be the optimal alignment, or a score that is equal to or greater than that generated by patentee's alignment, in which case the accused sequence still falls within the scope of the claim.

A critical issue when drafting a similarity score claim is the determination of an appropriate threshold similarity score (X in the above example). Of course, the score must be less than the theoretical maximum score (the score generated by aligning the sequence with itself) if it is to encompass any sequences beyond the recited sequence. At the same time, the score must be high enough to satisfy the requirements of novelty and nonobviousness, as well as the enablement and written description requirements.

With regard to novelty and nonobviousness, a similar sequence in the prior art might impose a lower limit on the threshold score, since the score would presumably have to be high enough to at least exclude the prior art molecule.[86] In principle, it should also be high enough to distinguish any variants that might be considered obvious in view of the prior art sequence. For example, a threshold that defined a genus so broadly that it would encompass single point mutants of a prior art sequence might be considered obvious in light of that prior art sequence. There is little guidance from the courts in this regard, and the determination would likely depend upon the specific sequences in question. The practitioner drafting a claim in this situation would have to use judgment in assessing the distance the genus needs to be from the closest prior art sequence. Of course, a series of dependent claims with increasing threshold scores could be used as insurance against a claim being found obvious for encompassing a genus that is too close to the prior art.

The other limitation on claim breadth is written description and enablement, in particular the requirement of some correlation between structure and predicted function. A determination has to be made as to how high the threshold score must be in order for it to be sufficiently predictive of homology and/or conserved function to

86. *See* 35 U.S.C. §§ 102–103 (2004).

support a valid claim. Note that while this determination is somewhat speculative, particularly in view of the scant guidance the courts have provided with respect to the required degree of predicitivity, the need to make such a determination is not new. At least implicitly, the same determination has to be made with percent identity claims, *i.e.*, the Patent Office will only allow percent identity limitations that it determines to be sufficiently predictive of homology.[87] The only difference is that similarity score is a more accurate predictor; for claimed sequence genuses of comparable size the validity of the similarity score is on much more solid ground.

One approach to selecting an appropriate threshold score would be to identify an allowable percent identity limitation and then use that number to calculate a corresponding similarity score. Take, for example, a case where one is claiming a particular amino acid sequence, and assume that an 80% identity limitation would be high enough to satisfy the enablement and written description requirement. One could simply calculate a similarity score that corresponds to an 80% identical sequence and use that similarity score as a claim limitation.

Of course, the similarity score can vary dramatically between different 80% identical pairings, depending upon the nature of the mismatches; conservative substitutions will yield relatively high similarity scores, while nonconservative substitutions and insertions/deletions will result in much lower similarity scores. This is illustrated in Examples 8 and 9, which depict conservative and nonconservative 80% identity alignments, respectively.

Example 8

```
RFWQRHDVHMIDFYYAWYQRHSVHCIDFAY
RFWQRHDIHMLDFYYSWYQRHAVHCLDFSY (171)24/30 80% identity
```

Both alignments share the same upper sequence, which would yield a theoretical maximum similarity score of 185 when aligned with itself.

Example 9

```
RFWQRHDVHMIDFYYAWYQRHSVHCIDFAY
RFNQRVDVCMIDFYYAKYQRHSVIPIDFAY (111)24/30 80% identity
```

87. *See Synopsis of Application of Written Description Guidelines*, *supra* note 41, at 46.

The conservative substitutions of Example 8 result in a score of 171, while nonconservative substitutions of Example 9 drop the score all the way down to 111. In fact, the score could have been substantially lower than 111 if gaps were introduced into the alignment. One drafting claims with an eye toward maximizing claim scope would prefer to recite a similarity score limitation corresponding to nonconservative substitutions, *e.g.*, 111. Logically, this limitation should be allowable since it is a biologically-based measure of the functional similarity of the 80% identical sequences, and we have assumed that the 80% identity limitation would have been acceptable under the current scheme. However, a similarity score limitation of 111 is much broader than an 80% identity limitation, since it would encompass alignments sharing much lower percent identity but more conservative substitutions. This is illustrated in Example 10, where the upper sequence from the Examples 8 and 9 alignments is aligned with a sequence sharing only 27% identity, but where the substitutions are conservative and as a result, the similarity score is relatively high at 115.

Example 10

```
RFWQRHDVHMIDFYYAWYQRHSVHCIDFAY
KYWEKHEIHLLNYFYSWFEQHAIHCLEYSF (115)8/30 27% identity
```

This nicely illustrates the ability of the similarity score approach to capture functionally similar sequences that are relatively distant in terms of percent identity. It also illustrates an inherent weakness to the percent identity approach, in that the biological significance of a given percent identity limitation will vary dramatically depending upon the nature of the mismatches.

Alternatively, instead of basing a similarity score claim limitation upon a corresponding percent identity claim limitation, an appropriate threshold similarity score could be arrived at *de novo* based upon the likelihood that it represents true homology and/or conserved function. As described in more detail below, the BLAST program can convert similarity scores into "expectation values" (*i.e.*, "E-values"), which are a measure of the likelihood that two similar sequences are truly homologous. Thus, one could use this type of calculation to assess how predictive a given threshold score is of homology for a particular reference sequence of interest, and draft claims accordingly.

Not only is the similarity score approach technically superior to percent identity, similarity scores are unambiguous and easily determined, either manually or with the aid of a computer. There is a perception among some that similarity scores (*e.g.*, BLAST scores) are generated by arcane and complex algorithms that can only be understood by one with a fairly sophisticated understanding of computer science or bioinformatics, and that the scores can only be practically calculated by means of a computer. As shown above, a similarity score for aligned sequences can be calculated manually with simple arithmetic, merely by applying a given scoring matrix to the aligned sequences (and if appropriate, employing a gap penalty). While the science underlying the generation of substitution matrices and the determination of appropriate gap penalties is fairly sophisticated, once the matrix and gap penalties have been defined, any person able to do simple arithmetic can readily calculate a score for a pair of aligned sequences of moderate length, *i.e.*, on the order of hundreds of amino acids, the length of a typical claimed protein sequence. Thus, it would be straightforward for a judge or jury to evaluate an alignment of the sequences recited in a claim and an allegedly infringing sequence and determine whether or not it falls within the literal scope of the claim. This enhances the transparency of claim interpretation, since one need not rely on a computerized "black box" to determine the score, and ultimately to determine whether an accused protein sequence literally infringes.

IV. USING THE BLAST PROGRAM WITH SIMILARITY SCORE CLAIMS

While it is advantageous that similarity scores can be calculated manually, in practice one would normally employ a computer to align and score similar sequences, *e.g.*, when drafting claims, assessing the validity of claims, or determining whether a particular sequence of interest falls within the scope of the claims. A number of computer programs are available that will fulfill this function, including BLAST, FAST-All (FASTA) and various implementations of the Smith-Waterman algorithm.[88] To illustrate, I will focus on the BLAST program at the NCBI website ("BLAST"), since it is freely

88. *See*, *e.g.*, NCBI BLAST, *supra* note 74; FASTA Protein Database Query, European Bioinformatics Institute, *at* http://www.ebi.ac.uk/fasta33/ (last visited Sept. 25, 2004); MPsrch Submission Form (Smith-Waterman algorithm), European Bioinformatics Institute, *at* http://www.ebi.ac.uk/MPsrch (last visited Sept. 25, 2004).

available via the Internet and configured to search the most extensive databases of protein sequences available in the public domain.[89]

A. An Overview of BLAST

When using BLAST to calculate a similarity score for two sequences, or to draft or analyze a similarity score claim, it is important that various adjustable parameters in the program be properly set. Bear in mind that the primary intended function of BLAST is to address biological questions, not to calculate similarity scores for the purpose of claim drafting. BLAST employs a variety of refinements to adjust a "raw" similarity score (*i.e.*, a score based only on substitution matrix and gap penalties), thereby fine-tuning the score to more accurately predict the evolutionary significance of the similarity between aligned sequences. While these refinements result in a more biologically relevant score, in the context of patent claiming they introduce unnecessary complexity, and in some cases ambiguity, to the score. In my view, this increased complexity and ambiguity is not justified by the relatively slight improvement in biological relevance. Therefore, my recommendation is not to use the refinements in drafting protein claims, but instead to use only the raw score.

For example, BLAST offers the option of using "composition-based statistics." In fact, it is used in the default setting of the program. Composition-based statistics employs a scaling procedure that in effect uses a slightly different scoring system for each sequence in the database being searched.[90] Because a different scoring system is used, a raw score (or "raw S score") obtained with this feature turned on will be slightly different than the raw score that would be obtained simply by using the scoring matrix and gap penalties. While from a biologist's perspective, composition-based scaling is of value since it will somewhat improve the accuracy of the score, it is undesirable in the context of patent claims since it renders

89. The Basic Local Alignment Search Tool (BLAST) program is commonly used by scientists to perform sequence alignments of protein sequences. The typical user is a researcher with a protein sequence of interest, *i.e.*, a query sequence. The scientist submits the query sequence and chooses a protein database, and the BLAST algorithm will search that database for similar sequences, based on similarity score. The output is a ranked listing of similar sequences exceeding some certain threshold score. Along with the similarity score, the program also provides alignments of all the sequences with the query sequence. Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well designed queries and alignments.

90. *See* NCBI BLAST, *supra* note 74.

the scores unstable. This instability arises because the scaling procedure takes into account all of the sequences in the database, and the sequence databases are highly dynamic, constantly changing as new sequence information is submitted by researchers.[91] As a result, scores obtained for identical sequences as calculated with composition-based statistics turned on will vary with time.[92] Clearly this is undesirable from a patenting standpoint, which requires a score that is stable and independent of the nature of any particular database. Turning off the composition-based statistics results in a stable score that is based solely on the scoring matrix and gap penalties.

Another default BLAST option that should be disabled when calculating a similarity score is the "low complexity filter." The low complexity filter masks off sections of the query sequence having low compositional complexity, *i.e.*, sections of the sequence that are predominantly made up of one or a few amino acids.[93] Common examples are acidic-, basic-, and proline-rich regions of a protein sequence. The filtering can eliminate statistically significant, but biologically uninteresting reports from the BLAST output, and hence, is a useful feature for most biologically relevant searches.[94] However, such filtering will alter the raw score by ignoring certain amino acid pairings. Therefore, in determining a similarity score for patent purposes, all filters, including the low complexity filter, should be disabled.[95]

It is also important when using BLAST to understand that each alignment generates several different scores, all ultimately derived from the raw alignment score. One needs to be able to distinguish between the scores, and identify the correct "raw score" for claiming purposes. For example, consider an excerpt from an actual BLAST output shown in Figure 2. Just above the alignment are two scores—a "raw alignment score," or "S" (in parenthesis), and a "bit score," or "S'". In this example, S is 952 and S' is 371 bits. The S score is the "similarity score" of interest to us, calculated based solely on scoring

91. *See* GenBank Overview, NCBI, *at* http://www.ncbi.nih.gov/Genbank/GenbankOverview.html (last visited Sept. 25, 2004).

92. This will be observed by anyone who does a BLAST search of a particular sequence (with composition based statistics turned on), keeps the results, and then re-runs the exact same search a month later. They will likely find that, as a result of this scaling feature and additions to the database over the course of the month, the raw score for alignment of the exact same sequences will have changed.

93. *See* NCBI BLAST, *supra* note 74.

94. *See id*.

95. There are a number of other optional settings. For detailed explanations of all these settings, see NCBI BLAST, *supra* note 74.

matrix and gap penalty. The bit score (S') is a normalized score derived from the raw alignment score S in a manner which takes into account the statistical properties of the scoring system.[96] Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches that were based on different scoring systems. Thus, for the biologist, the advantage of using the bit score instead of the raw score comes when one is comparing the biological relevance of scores obtained using a different scoring system (*i.e.*, a different scoring matrix and/or gap penalty). This is not an issue for patent claiming purposes. The disadvantage of using a bit score for claiming purposes is that it introduces unnecessary complexity and ambiguity into the score.[97]

96. To convert a raw score S into a normalized score S' expressed in bits, one uses the formula S' = (lambda*S - ln K)/(ln 2), where lambda and K are parameters dependent upon the scoring system (substitution matrix and gap costs) employed. For a detailed explanation of bit scores, see Lambda Ratio, NCBI, *at* http://www.ncbi.nlm.nih.gov/BLAST/matrix_info.html#lambda (last visited May 25, 2004).

97. *Id*.

Figure 2[98]

```
>gi|3913985|sp|Q44287|LAMB_AERSA   Maltoporin precursor (Maltose-inducible porin)
 gi|541232|pir||S37779   porin precursor, maltose-inducible - Aeromonas salmonicida
 gi|398210|emb|CAA49223.1|   maltose-inducible porin [Aeromonas salmonicida]
          Length = 445

 Score =  371 bits (952), Expect = e-101
 Identities = 200/447 (44%), Positives = 271/447 (60%), Gaps = 32/447 (7%)

Query: 1   MKTSLRTLSVALAAALVSPSVLAIEKIDFHGYMRAGVGVSSDG----GLAEWQKTMVGRL 56
            MK     ++  + AAL S +  A+   DFHGY R+GVGVS+DG    GL++  K  VGRL
Sbjct: 3   MKAKWLPIAAGVTAALASQAAFAV---DFHGYFRSGVGVSTDGSMQTGLSDNAKQKVGRL 59

Query: 57  GNESDTYGEIGLGAEVYKKEDVSFYLDSMVSMLSDGSNDSETTIG---------------- 101
            GNE+DTYGEI LG+EV+ K+  +FY+DSMV+M S+GSND E+T
Sbjct: 60  GNEADTYGEIQLGSEVFNKDGKTFYVDSMVAMTSNGSNDWESTESKFQCTSANGTALDGC 119

Query: 102 ---DDAQFGLRQLNLQIKGLIPGDKEAVIWGGKRYYQRHDLHIIDTKYWNISGSGAGIEN 158
               +DA F LRQ N+Q KGL+    EA +W GKRYYQRHD+HI D  YWNISG GAGIE
Sbjct: 120 ENKEDATFALRQFNVQAKGLLGFAPEATLWAGKRYYQRHDVHISDFYYWNISGRGAGIEG 179

Query: 159 YTVGPGAVSVAWVRGDANDVDTRITGDSDVNINYIDVRYAGFKPWAGSWTEVGIDYAMPN 218
               GPG VS AWVR D +  +   T + ++N+N +D+RYAG   W     EVG+DYA+ N
Sbjct: 180 IQAGPGKVSFAWVRNDRSGTNVDGTYNDEMNVNTLDLRYAGIPLWQDGSLEVGVDYAIAN 239

Query: 219 PTKQQKEYGGLY--DADNAVMLTGEISQDMFGGYNKLVLQYANKGLAQNMISQGG-GWYD 275
           P+  QK+       +A + VMLT E++Q + GG+NK VLQY  +G ++     G   WY
Sbjct: 240 PSDAQKDSANAQYKNAKDGVMLTAELTQGILGGFNKTVLQYGTEGYSKTFAFWGDRSWYG 299

Query: 276 MWHKTDEAKGYRVINTGLIPITDKFSFNHVLTWGSANDITEYTDKTNLISLVGRAQYQFT 335
               K D A G+R+IN G+IP+ + +   H L +G  ND+ +  DK   +S+V R  Y++
Sbjct: 300 AEAK-DGADGFRIINHGVIPMGNSWEMGHQLVYGVGNDMWDTNDKWETMSVVARPMYKWD 358

Query: 336 QYVRAIGEVGGFYQKDTYHNGSNYKQGGEKYTIALGLAEGPDFLSRPELRVFASYLNDSE 395
            + + I E G F  K+   NG++ +  G K T+A   + G  F +RPE+RVFASYL   +
Sbjct: 359 DFNKTIFEGGYFKDKNKSTNGTSEEDAGYKLTLAQAWSAGSSFWARPEIRVFASYLAQDK 418

Query: 396 ---NGKPFEDGTSNDTWNFGVQVEAWW 419
               G  F +GT++DTWNFGVQ EAWW
Sbjct: 419 KEMKGNAFNNGTADDTWNFGVQAEAWW 445
```

98. Figure 2 is an excerpt from the output of a BLASTP search conducted on the NCBI website, *supra* note 74, on May 25, 2004.

Along with the raw and bit scores, the "expectation," or "E value" is provided. For example, in Figure 2 the E value is e^{-101}. The E value is a statistical measure derived from the score, and represents the number of different alignments with scores equivalent to or better than S that are expected to occur by chance in a search of a sequence of the size of the query sequence in the database searched.[99] The significance of this value is that in a database containing a large number of sequences, by random chance there are going to be some sequences that have a certain degree of similarity with the query sequence. The higher the S score of an alignment, however, the less likely it is that the alignment is the result of chance similarity and the more likely that it represents true homology. The formula used by BLAST to calculate E is $E=mN2^{-S'}$, where S' is the bit score and "m" and "N" are, respectively, the lengths of the query sequences and the total length of the database in residues.[100] The lower the E value, the more significant the score, *i.e.*, the higher the likelihood that this indeed represents an alignment of homologous sequences as opposed to a chance similarity appearing in two sequences of unrelated origin.[101]

The E value is typically the score that a biologist reviewing a BLAST output would be most interested in, since it provides the best intuitive measure of how close the aligned sequences are to one another, *e.g.*, an E value of 10^{-6} represents that there is only a one in a million chance that an alignment with that high of a score would be achieved for the given query sequence and database searched, *i.e.*, there is a very high degree of confidence that this alignment reflects true homology. On the other hand, an E value of ten represents that just by chance ten alignments with a score of that magnitude would be expected to be found by the search. In this case, there is much less confidence that the aligned sequences are actually related.

While the E value provides the best measure of the likelihood of homology between two sequences, it is not appropriate for use as a substitute for percent identity in claiming proteins. Not only does it have the ambiguity and undue complexity associated with S', but the

99. Samuel Karlin & Stephen F. Altschul, *Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes*, 87 PROC. NAT'L ACAD. SCI. USA 2264 (1990), *available at* http://www.pnas.org/cgi/reprint/87/6/2264.pdf (last visited Oct. 25, 2004).

100. The database is treated as a single long sequence of N residues. Note that the E value incorporates the raw score (S), the statistical nature of the scoring system (K and lambda), and the size of the query and database (m and N).

101. *See* NCBI BLAST, *supra* note 74.

value of E will vary dramatically depending upon the size of the database searched (N in the equation). Thus, the value only has meaning in the context of a search of a particular database. BLAST offers the option of searching any of six different databases, each of which has a different size, depending upon the number of sequences in the database and the length of the sequences.[102] The larger the database searched, the larger the value of E, and hence the greater the likelihood that the scoring alignment is purely the result of chance.

B. Determining Infringement of Similarity Score Claims

Suppose one would like to know whether a sequence of interest (a "query sequence") falls within the scope of a similarity score claim. It would be necessary to determine the similarity score for an optimal alignment of the claim's reference sequence and the query sequence. A simple way to calculate this similarity score would be to use the BLAST 2 Sequences tool provided on the NCBI website.[103] To do so, a user simply inputs the reference and query sequences as the sequences to be scored (*e.g.*, by pasting in the sequences, with the amino acids represented by their single letter symbols), chooses the desired scoring matrix and gap penalties (those suggested herein, *i.e.*, the BLOSUM62 matrix and gap existence and extension penalties of 11 and 1 are the current defaults), turns off all filters, and clicks the alignment button. The program will optimally align the sequences and provide the similarity score. As described above, BLAST 2 reports three scores, the raw S score being the similarity score. For example, if the sequences of Example 4 are inputted, the score is reported as "Score = 32.3 bits (72), Expect = 2.3." The similarity score (raw score) is in parentheses (72). If the similarity score exceeds that recited in the claim, the query sequence falls within the literal scope of the claim.

C. Examining or Analyzing Similarity Score Claims

As another example, a patent examiner could use BLAST to analyze a similarity score claim with respect to the prior art.[104] The NCBI BLAST tool allows one to search what is referred to as the "nr" database. This is a non-redundant protein sequence database

102. *See id*.

103. *See* NCBI BLAST2 Sequences, NCBI, *at* http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html (last visited May 25, 2004).

104. 35 U.S.C. § 131 (2004) provides the statutory authority for the PTO to examine patent applications.

compiled from a variety of sources, and is the largest and most complete single database of protein sequences in the public domain.[105]

In particular, to run a BLAST search on a reference sequence to determine if any public domain sequences fall within the similarity score limitation, one can access the BLAST program at the NCBI website.[106] The "protein-protein BLAST ("blastp")" should then be selected, which will pull up the protein-protein BLAST page (the most basic version of BLAST for comparing polypeptide sequences). There the user is presented with a number of fields to be completed. This discussion will focus only on those fields that are relevant for our purposes; many will simply be left at their default setting. One important field is "Search"; this is where the claim's reference sequences (the sequence can simply be copied and pasted in) should be entered.

Next, the proper database to search must be chosen. There are currently six databases to choose from; we will stay with the default "nr" database, as described above.

Moving down to the "Options for Advanced Blasting" section of the page, the "composition-based statistics" and the "low complexity filter," should be turned off. These are on by default, but for the reasons previously discussed, should be disabled when determining a similarity score.

The program allows the user to choose the scoring matrix and gap costs. We will use the default settings of BLOSUM62, gap existence penalty of 11 and gap extension penalty of 1. Of course, one could decide to use a different matrix and/or gap penalties in a similarity score claim, and in that case, these settings should be adjusted.

Other parameters (*e.g.*, the threshold level for expectation values, etc.) will not affect how the search algorithm is implemented, and should normally be left in their default settings.

As an example, consider an artificial sequence concocted by taking the sequence of maltoporin precursor protein and randomly changing 102 of its amino acid residues.[107] The full length of the sequence is 419 amino acids. Therefore, this artificial mutant retains

105. There is no single database that contains all sequences that might be considered prior art, but the nr database, or for that matter any database that the PTO would like to search, can be searched using the BLAST tool.

106. *See* NCBI BLAST, *supra* note 74.

107. *See supra* note 98. This sequence is not intended to have any biological relevance, but is used merely to exemplify the mechanical operation of the BLAST tool and its application to the similarity score approach to claiming proteins.

about 75% identity to the original sequence. To run a BLAST search, access the blastp page, paste the sequence into the search field, select the nr database and adjust the default settings as indicated above, and then submit the query. After the query has run, a BLAST output is given.

Excerpts from the BLAST output obtained for this query on May 25, 2004 are presented in Figures 3 and 4. Figure 3 shows the ranked listing of top scoring related sequences. The top scoring sequence (gi|16121158|ref|NP_404471.1|) is the native maltoporin precursor sequence. Thus, even after changing 25% of the amino acid positions, the most similar sequence is still the starting sequence.

Figure 3

Related Structures

Sequences producing significant alignments:	Score (bits)	E Value
gi|16121158|ref|NP_404471.1| maltoporin [Yersinia pestis] >...	627	e-178
gi|22127111|ref|NP_670534.1| putative outer membrane protei...	627	e-178
gi|45443284|ref|NP_994823.1| maltoporin [Yersinia pestis bi...	624	e-177
gi|3913985|sp|Q44287|LAMB_AERSA Maltoporin precursor (Malto...	272	1e-71
gi|398211|emb|CAA49224.1| maltose-inducible porin [Aeromona...	271	1e-71
gi|31340212|sp|Q8KKH1|LMB2_AERHY Maltoporin precursor (Malt...	263	5e-69
gi|31340211|sp|Q8KKH0|LMB1_AERHY Maltoporin precursor (Malt...	261	2e-68
gi|30793638|gb|AAP40342.1| Omp48 protein precursor [Aeromon...	254	2e-66
gi|28901499|ref|NP_801154.1| maltose-inducible porin [Vibri...	245	1e-63
gi|46913859|emb|CAG20641.1| hypothetical maltoporin [Photob...	226	9e-58
gi|15601781|ref|NP_233412.1| maltoporin [Vibrio cholerae O1...	148	2e-34
gi|3914229|sp|Q56652|LAMB_VIBCH Maltoporin precursor (Malto...	146	9e-34
gi|32035629|ref|ZP_00135540.1| COG4580: Maltoporin (phage l...	137	3e-31
gi|16123849|ref|NP_407162.1| maltoporin [Yersinia pestis] >...	132	1e-29
gi|45442828|ref|NP_994367.1| maltoporin [Yersinia pestis bi...	132	1e-29
gi|22123953|ref|NP_667376.1| maltose high-affinity receptor...	132	1e-29
gi|280123|pir||A60177 LamB maltoporin protein precursor - S...	131	3e-29
gi|16767481|ref|NP_463096.1| maltoporin precursor [Salmonel...	130	4e-29
gi|16762912|ref|NP_458529.1| maltoporin precursor [Salmonel...	130	7e-29
gi|2098396|pdb|2MPR|A Chain A, Maltoporin From Salmonella T...	129	9e-29
gi|1941972|pdb|1MPR|A Chain A, Maltoporin From Salmonella T...	128	3e-28
gi|24115373|ref|NP_709883.1| phage lambda receptor protein;...	127	6e-28
gi|16131862|ref|NP_418460.1| phage lambda receptor protein ...	127	6e-28
gi|30064627|ref|NP_838798.1| maltose high-affinity receptor...	127	6e-28
gi|15804629|ref|NP_290670.1| phage lambda receptor protein;...	127	6e-28
gi|26250818|ref|NP_756858.1| Maltoporin precursor [Escheric...	127	6e-28
gi|396371|gb|AAC43130.1| phage lambda receptor protein	125	1e-27
gi|400158|sp|P31242|LAMB_KLEPN Maltoporin precursor (Maltos...	125	2e-27

Figure 4

Alignments

```
>gi|16121158|ref|NP_404471.1|   maltoporin [Yersinia pestis]
 gi|20138636|sp|Q8ZHP0|LMB2_YERPE    Maltoporin precursor 2 (Maltose-inducible
porin)
 gi|25298933|pir||AF0104     maltoporin [imported] - Yersinia pestis (strain
CO92)
 gi|15978924|emb|CAC89697.1|   maltoporin [Yersinia pestis CO92]
          Length = 419

 Score =  627 bits (1616), Expect = e-178
 Identities = 317/419 (75%), Positives = 339/419 (80%)

Query: 1   MKVSLKTGTLVAAASLVGPSGPAIDKHIFHMYARSLITVCKDGGLAEWDKTMVERLGNES 60
           MK SL+T ++  AA+LV PS  AI+K  FH Y R+ + V  DGGLAEW KTMV RLGNES
Sbjct: 1   MKTSLRTLSVALAAAALVSPSVLAIEKIDFHGYMRAGVGVSSDGGLAEWQKTMVGRLGNES 60

Query: 61  DTYGFIHLGAEKYKQHDVSYYGDSMVSTLGDGSNDSPWTIGNQAQFGLRQLNLQPKGEIP 120
           DTYG I LGAE YK+ DVS+Y DSMVS L DGSNDS  TIG+ AQFGLRQLNLQ KG IP
Sbjct: 61  DTYGEIGLGAEVYKKEDVSFYLDSMVSMLSDGSNDSETTIGDDAQFGLRQLNLQIKGLIP 120

Query: 121 GDKEAVRSGGSRYYQRHDLHILCTKYWNISGSGAGIENYTVGPGAVSVAWVRGDANDVDT 180
           GDKEAV  GG RYYQRHDLHI+ TKYWNISGSGAGIENYTVGPGAVSVAWVRGDANDVDT
Sbjct: 121 GDKEAVIWGGKRYYQRHDLHIIDTKYWNISGSGAGIENYTVGPGAVSVAWVRGDANDVDT 180

Query: 181 RITGDSDVNINYVDVRYAGFCSWAGSWTEVGIDYAMIAGHKQMKVYGEQFDVINAVMLTG 240
           RITGDSDVNINY+DVRYAGF  WAGSWTEVGIDYAM    KQ K YG  +D  NAVMLTG
Sbjct: 181 RITGDSDVNINYIDVRYAGFKPWAGSWTEVGIDYAMPNPTKQQKEYGGLYDADNAVMLTG 240

Query: 241 TISQDMFGNYNKDSTQYANKGLAQPEIQQGGAWYDMVHKHDNPKGWRVLNTGLFPASDKF 300
            ISQDMFG YNK   QYANKGLAQ  I QGG WYDM HK D  KG+RV+NTGL P +DKF
Sbjct: 241 EISQDMFGGYNKLVLQYANKGLAQNMISQGGGWYDMWHKTDEAKGYRVINTGLIPITDKF 300

Query: 301 TFNHVLTGWEENDITEYQDKVQVISLVGRMQYQFSQYVRAIGEVGGFYQKDTYSNLSNFI 360
           +FNHVLT    NDITEY DK  +ISLVGR QYQF+QYVRAIGEVGGFYQKDTY N SN+
Sbjct: 301 SFNHVLTWGSANDITEYTDKTNLISLVGRAQYQFTQYVRAIGEVGGFYQKDTYHNGSNYK 360

Query: 361 NAGEKYTIALGLAEGMDFTSRPELRVFASYLQESENGYGFEMGTSNQTWNFGVQVESWW 419
            GEKYTIALGLAEG DF SRPELRVFASYL +SENG  FE GTSN TWNFGVQVE+WW
Sbjct: 361 QGGEKYTIALGLAEGPDFLSRPELRVFASYLNDSENGKPFEDGTSNDTWNFGVQVEAWW 419
```

Figure 4 is the alignment between the query sequence and native maltoporin precursor. The similarity score is 1,616, *i.e.*, the "raw score" appearing in parenthesis after the score in bits. Hence, the native maltoporin precursor sequence would be encompassed by a claim reciting this query sequence as a reference sequence and a similarity score limitation less than or equal to 1,616. If that were the case, the examiner would need to investigate this and any other sequences that fall within the limitation to determine whether they are prior art, *e.g.*, whether the putative prior art sequence was actually published prior to the invention of the claimed sequence.[108] If the sequence is in the prior art and there are no other claim limitations to distinguish it, *e.g.*, a functional limitation that excludes the native maltoporin precursor sequence, then the examiner would generally reject the claim as anticipated. The applicant might then respond by amending the claims to increase the similarity score threshold to avoid the prior art sequences.

This same sort of analysis could also be conducted to assess the validity of issued claims. For example, a third-party interested in evaluating the validity of the claim could run the same BLAST search post-issuance. If a sequence that falls within the recited similarity score range is found to exist in the prior art, this might form the basis of an argument that the claim is anticipated, invalid and should not have been issued by the Patent Office. This analysis, of course, might depend upon a determination of whether the prior art sequence possesses any functional limitation recited in the claims.

D. Drafting Similarity Score Claims

Suppose that an inventor has discovered what is believed to be a novel, patentable protein sequence and wishes to use it as a basis for a patent claim directed to a genus of similar protein sequences, using the similarity score approach. A first step might be to submit the sequence as a BLAST query against the nr database in order to (1) confirm that the sequence is indeed novel, (2) identify the closest similar sequences in the public domain, and (3) to obtain similarity score for the alignment of the query sequence with similar public domain sequences.

108. This information can usually be gleaned from the GenBank record for the sequence, which contains information regarding the date when the sequence was first posted, publication information, etc. Conveniently, the GenBank records of sequences are normally hyperlinked to the identifier as it appears in the BLAST output. *See* GenBank Overview, *supra* note 91.

To illustrate, let us consider the same artificial sequence analyzed above, but this time imagine it as a novel sequence one is interested in patenting. Again, we would find that the top scoring sequence is the native maltoporin precursor sequence from which the query sequence was derived. The similarity score, *i.e.*, the "raw score," is 1,616. Depending upon the rigor with which we wish to search the prior art, we might decide to try the BLAST against other databases to which we might have access. For the purposes of this illustration, we will assume we are satisfied with the result of searching the nr database.

Because the sequence appears to be novel, with no closely similar molecules appearing in the public domain, it appears that prior art is not a bar to patenting the molecule. In this scenario, we have identified the most similar sequence as being the maltoporin precursor. In light of the relatively high similarity between the sequences, we might be able to infer something about the function of our novel sequence, *i.e.*, if the novel sequence was isolated from a natural source, the high similarity score with maltoporin precursor indicates likely homology and conserved function.

Of course, the scope of a genus claim encompassing the sequence can be limited by the proximity of the prior art. In this case, the closest prior art sequence has a raw score of 1,616. Thus, any claimed genus defined using the novel sequence as the reference sequence and a similarity score limitation would probably need to recite a score higher than 1,616 to avoid reading on the prior art sequence.

In determining the magnitude of the similarity score threshold for the claim, it is useful to know the theoretical maximum score for an alignment with the novel sequence. This would be the score for alignment of the sequence with the identical sequence, *i.e.*, 100% identity. A convenient way to determine this score is by means of the BLAST 2 Sequences (BL2S) tool described above, *i.e.*, by simply aligning the novel sequence with itself. In this example, the calculated raw score S is 2,258, which represents the maximum score for any alignment with this sequence.

In drafting a claim, the inventor in this scenario would generally base his claim on a similarity score limitation lying somewhere between 1,616 (the score of the closest prior art sequence) and 2,258 (the theoretical maximum score). A score of 1,617 would provide optimal scope of coverage from the patentee's point of view, but it is questionable whether the Patent Office or the courts would allow such broad coverage. The claim would presumably not be anticipated, but

might be found obvious for claiming a large number of sequences that are very close to the prior art sequence.[109] The value ultimately allowed by the PTO should lie somewhere between 1,617 and 2,258, and would need to be sufficiently high to satisfy the requirements of § 112. As discussed above, an appropriate threshold value can be derived from an allowable percent identity limitation, and/or could be informed by converting a proposed similarity score limitation to a corresponding E value to assess the number's biological significance.

V. CONCLUSION

For a variety of reasons, the similarity score approach represents a more rational and scientific basis to claiming a genus of related proteins compared to the current practice based on percent identity. While this approach might meet initial resistance from some patent examiners or practitioners unfamiliar with the concept of similarity score, by educating these individuals to the advantages of similarity score it should be possible to convince them to accept this as a valid, and indeed superior, alternative to percent identity claiming. The widespread adoption of this approach would result in more effective claim coverage for the patentee, with a greater likelihood that the claims will withstand a challenge to validity during litigation.

109. For example, a claim reciting a threshold score of 1,617 would encompass many point mutants of the maltorporin precursor sequence, including any mutant where any one of the 102 amino acid differences is substituted by the corresponding residue in the claimed reference sequence.