



(86) Date de dépôt PCT/PCT Filing Date: 2005/03/22  
(87) Date publication PCT/PCT Publication Date: 2005/10/06  
(45) Date de délivrance/Issue Date: 2013/01/15  
(85) Entrée phase nationale/National Entry: 2006/09/21  
(86) N° demande PCT/PCT Application No.: US 2005/009557  
(87) N° publication PCT/PCT Publication No.: 2005/092059  
(30) Priorité/Priority: 2004/03/22 (US60/555,520)

(51) Cl.Int./Int.Cl. *C12Q 1/68* (2006.01),  
*C12N 15/11* (2006.01), *C12N 15/33* (2006.01),  
*C12N 15/40* (2006.01), *C12P 19/34* (2006.01)

(72) Inventeurs/Inventors:  
SAMPATH, RANGARAJAN, US;  
HALL, THOMAS A., US;  
ESHOO, MARK W., US

(73) Propriétaire/Owner:  
IBIS BIOSCIENCES, INC., US

(74) Agent: SMART & BIGGAR

(54) Titre : COMPOSITIONS POUR L'IDENTIFICATION DE LA FIEVRE HEMORRAGIQUE VIRALE  
(54) Title: COMPOSITIONS FOR USE IN IDENTIFICATION OF VIRAL HEMORRHAGIC FEVER VIRUSES

(57) **Abrégé/Abstract:**

The present invention provides oligonucleotide primers, compositions, and kits containing the same for rapid identification of viruses that cause viral hemorrhagic fevers by amplification of a segment of viral nucleic acid followed by molecular mass analysis.

## (12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property  
Organization  
International Bureau



(43) International Publication Date  
6 October 2005 (06.10.2005)

PCT

(10) International Publication Number  
**WO 2005/092059 A3**

(51) International Patent Classification<sup>7</sup>: **C12Q 1/68**,  
C12P 19/34, C12N 15/11, 15/33, 15/40

(21) International Application Number:  
PCT/US2005/009557

(22) International Filing Date: 22 March 2005 (22.03.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/555,520 22 March 2004 (22.03.2004) US

(71) Applicant (for all designated States except US): **ISIS  
PHARMACEUTICALS, INC.** [US/US]; 2292 Faraday  
Avenue, Carlsbad, CA 92008 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **SAMPATH, Ran-  
garajan** [US/US]; 12223 Mannix Road, San Diego, CA  
92129 (US). **HALL, Thomas, A.** [US/US]; 5239 Wohlford  
Street, Oceanside, CA 92056 (US). **ESHOO, Mark, W.**  
[US/US]; 615 Glenmont Drive, Solana Beach, CA 92075  
(US).

(74) Agent: **LEGAARD, Paul, K.**; Cozen O'Connor, 1900  
Market Street, Philadelphia, PA 19103 (US).

(81) Designated States (unless otherwise indicated, for every  
kind of national protection available): AE, AG, AL, AM,  
AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN,  
CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI,  
GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE,  
KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD,  
MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG,  
PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ,  
TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA,  
ZM, ZW.

(84) Designated States (unless otherwise indicated, for every  
kind of regional protection available): ARIPO (BW, GH,  
GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,  
ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),  
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,  
FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO,  
SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN,  
GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

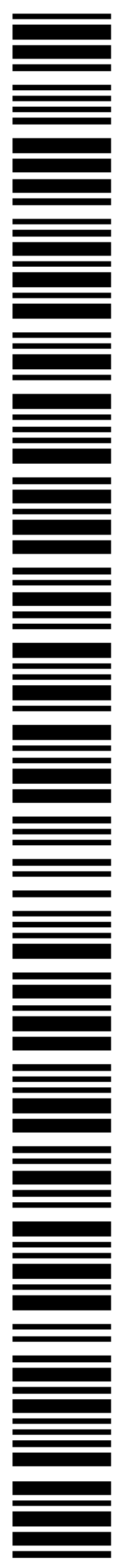
- with international search report
- before the expiration of the time limit for amending the  
claims and to be republished in the event of receipt of  
amendments

(88) Date of publication of the international search report:  
1 December 2005

For two-letter codes and other abbreviations, refer to the "Guid-  
ance Notes on Codes and Abbreviations" appearing at the begin-  
ning of each regular issue of the PCT Gazette.

(54) Title: COMPOSITIONS FOR USE IN IDENTIFICATION OF VIRAL HEMORRHAGIC FEVER VIRUSES

(57) Abstract: The present invention provides oligonucleotide primers, compositions, and kits containing the same for rapid iden-  
tification of viruses that cause viral hemorrhagic fevers by amplification of a segment of viral nucleic acid followed by molecular  
mass analysis.



**WO 2005/092059 A3**

# **DEMANDES OU BREVETS VOLUMINEUX**

**LA PRÉSENTE PARTIE DE CETTE DEMANDE OU CE BREVETS  
COMPREND PLUS D'UN TOME.**

**CECI EST LE TOME \_\_1\_\_ DE \_\_2\_\_**

NOTE: Pour les tomes additionels, veuillez contacter le Bureau Canadien des Brevets.

---

# **JUMBO APPLICATIONS / PATENTS**

**THIS SECTION OF THE APPLICATION / PATENT CONTAINS MORE  
THAN ONE VOLUME.**

**THIS IS VOLUME \_\_1\_\_ OF \_\_2\_\_**

NOTE: For additional volumes please contact the Canadian Patent Office.

## COMPOSITIONS FOR USE IN IDENTIFICATION OF VIRAL HEMORRHAGIC FEVER VIRUSES

### STATEMENT OF GOVERNMENT SUPPORT

[0002] This invention was made with United States Government support under DARPA/SPO contract BAA00-09. The United States Government may have certain rights in the invention.

### FIELD OF THE INVENTION

[0003] The present invention relates generally to the field of genetic identification and quantification of viruses in the Filoviridae, Flaviviridae, Bunyaviridae and Arenaviridae families and provides methods, compositions and kits useful for this purpose, as well as others, when combined with molecular mass analysis.

### BACKGROUND OF THE INVENTION

#### A. Viral Hemorrhagic Fever

[0004] Viral hemorrhagic fevers (VHFs) are a group of febrile illnesses caused by RNA viruses from several viral families. These highly infectious viruses lead to a potentially lethal disease syndrome characterized by fever, malaise, vomiting, mucosal and gastrointestinal (GI) bleeding, edema and hypotension. The four viral families known to cause VHF disease in humans include Arenaviridae, Bunyaviridae, Filoviridae and Flaviviridae.

[0005] In acute VHF, patients are extremely viremic, and mRNA evidence of multiple events cytokine activation exists. *In vitro* studies reveal these cytokines lead to shock and increased vascular permeability, the basic pathophysiologic processes most often seen with VHF. Multi-system organ failure affecting the hematopoietic, neurologic and pulmonary systems often accompanies the vascular involvement. Another prominent pathologic feature is pronounced macrophage involvement. Inadequate or delayed immune response to these novel viral antigens may lead to rapid development of overwhelming viremia. Extensive infection and necrosis of affected organs also are described. Hemorrhagic complications are multifactorial and are related to hepatic damage, consumptive coagulopathy and primary marrow injury to megakaryocytes.

Aerosol transmission of some VHF viruses is reported among nonhuman primates and likely is a mode of transmission in patients with severe infection. Specific symptoms of VHF and modes of transmission vary depending on the particular viral pathogen.

## **B. Filoviruses**

[0006] Filoviruses are enveloped viruses with a genome consisting of one linear single-stranded RNA segment of negative polarity. The viral genome encodes 7 proteins.

Nucleoprotein (NP), virion protein 35 kDa (VP35) and virion protein 30 kDa (VP30) are associated with the viral ribonucleoprotein complex. VP35 is known to be required for virus replication and is thought to function as a polymerase cofactor. The viral RNA-dependent RNA polymerase is termed L (for large protein). The matrix protein (VP40) is the major protein of the viral capsid. The remaining proteins include virion glycoprotein (GP) and membrane-associated protein (VP24), which is thought to form ion channels. The Ebola viruses have one additional protein, small secreted glycoprotein (SGP).

[0007] Members of the filovirus genus include Zaire Ebola virus, Sudan Ebola virus, Reston Ebola virus, Cote d'Ivoire Ebola virus and Marburg virus. Ebola and Marburg viruses can cause severe hemorrhagic fever and have a high mortality rate. Ebola virus (Zaire and Sudan species) was first described in 1976 after outbreaks of a febrile, rapidly fatal hemorrhagic illness were reported along the Ebola River in Zaire (now the Democratic Republic of the Congo) and Sudan. Sporadic outbreaks have continued since that time, usually in isolated areas of central Africa. In 1995, eighteen years after the first outbreak was reported, Zaire Ebola reemerged in Kikwit, Zaire with 317 confirmed cases and an 81% mortality rate. The natural host for Ebola viruses is still unknown. Marburg virus, named after the German town where it was first reported in 1967, is primarily found in equatorial Africa. The host range of Marburg virus includes non-human and human primates. Marburg made its first appearance in Zimbabwe in 1975 and was later identified in other African countries, including Kenya (1980 & 1987) and Democratic Republic of the Congo (1999). Marburg hemorrhagic fever is characterized by fever, abdominal pain, hemorrhage, shock and a mortality rate of 25% or greater ("The Springer Index of Viruses," pgs. 296-303, Tidona and Darai eds., 2001, Springer, New York).

## **C. Flaviviruses**

[0008] Flaviviridae is a family of viruses that includes the genera flavivirus, hepacivirus and pestivirus. Viruses in the genus flavivirus are known to cause VHFs. Flaviviruses are enveloped

viruses with a genome consisting of one linear single-stranded RNA segment of positive polarity. The RNA genome has a single open reading frame and is translated as a polyprotein. The polyprotein is co- and post-transcriptionally cleaved by cell signal peptidase and the viral protease to generate individual viral proteins. Viral structural proteins include capsid (C), precursor to M (prM), minor envelope (M) and major envelope (E). Flavivirus non-structural proteins include NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5. NS1, NS2A, NS3 and NS4A are found in the viral replicase complex. In addition, NS3 is known to function as the viral protease, helicase and NTPase. NS2B is a co-factor for the protease function of NS3. NS5 is the viral RNA-dependent RNA polymerase and also has methyltransferase activity.

**[0009]** Members of the flavivirus genus include yellow fever virus, Apoi virus, Aroa virus, Bagaza virus, Banzi virus, Bouboui virus, Bukalasa bat virus, Cacipacore virus, Carey Island virus, Cowbone Ridge virus, Dakar bat virus, dengue virus, Edge Hill virus, Entebbe bat virus, Gadgets Gully virus, Ilheus virus, Israel turkey meningoencephalomyelitis virus, Japanese encephalitis virus, Jugra virus, Jutiapa virus, Kadam virus, Kedougou virus, Kokobera virus, Koutango virus, Kyasanur Forest disease virus, Langat virus, Louping ill virus, Meaban virus, Modoc virus, Montana myotis leukoencephalitis virus, Murray Valley encephalitis virus, Ntaya virus, Omsk hemorrhagic fever virus, Phnom Phenh bat virus, Powassan virus, Rio Bravo virus, Royal Farm virus, Saboya virus, Sal Vieja virus, San Perlita virus, Saumarez Reef virus, Sepik virus, St. Louis encephalitis virus, Tembusu virus, tick-borne encephalitis virus, Tyuleni virus, Uganda S virus, Usutu virus, Wesselsbron virus, West Nile virus, Yaounde virus, Yokose virus, Zika virus, cell fusing agent virus and Tamana bat virus.

**[0010]** A number of flaviviruses cause human disease, particularly hemorrhagic fevers and encephalitis. Each species of flavivirus has a unique geographic distribution; however, taken together, flaviviruses, and flavivirus-induced disease, can be found world-wide. One of the more commonly known diseases is dengue fever, or dengue hemorrhagic fever/shock, which was first described as a virus-induced illness in 1960. Dengue fever occurs in tropical and temperate climates and is spread by *Aedes* mosquitoes. The mortality rate is 1-10% and symptoms include febrile headache, joint pain, rash, capillary leakage, hemorrhage and shock. Another common flavivirus-induced disease is yellow fever. Yellow fever is found in tropical Africa and America and is transmitted by mosquitoes. The mortality rate is approximately 30% and symptoms include febrile headache, myalgia (muscle pain), vomiting and jaundice. Examples of some of the other diseases caused by flavivirus species include Japanese encephalitis, Kyasanur Forest

disease, Murray Valley encephalitis, Omsk hemorrhagic fever, St. Louis encephalitis and West Nile fever. The mortality rate of these diseases ranges from 0-20%. These diseases share many of the same symptoms, which may include headache, myalgia, fever, hemorrhage, encephalitis, paralysis and rash ("The Springer Index of Viruses," pgs. 306-319, Tidona and Darai eds., 2001, Springer, New York).

#### **D. Bunyaviridae**

[0011] Bunyaviridae is a family of viruses that includes the genera bunyavirus, phlebovirus, nairovirus, hantavirus and tospovirus. Viruses in three of these genera, hantavirus, phlebovirus and nairovirus, are known to cause VHF. Members of the Bunyaviridae family are enveloped viruses with a genome that consists of 3 single-stranded RNA segments of negative polarity. The genome segments are designated S (small), M (medium) and L (large). The S segment encodes the nucleocapsid protein (N). The two viral glycoproteins (G1 and G2) are encoded by the M segment and the L segment encodes the viral RNA-dependent RNA polymerase (L). For some Bunyaviridae species, additional viral non-structural proteins are encoded by the S and/or M segment ("The Springer Index of Viruses," pgs. 141-174, Tidona and Darai eds., 2001, Springer, New York).

[0012] Members of the hantavirus genus include, Hantaan virus, Seoul virus, Dobrava-Belgrade virus, Thailand virus, Puumala virus, Prospect Hill virus, Tula virus, Khabarovsk virus, Topografov virus, Isla Vista virus, Sin Nombre virus, New York virus, Black Creek virus, Bayou virus, Caño Delgadito virus, Rio Mamore virus, Laguna Negra virus, Muleshoe virus, El Moro Canyon virus, Rio Segundo virus, Andes virus and Thottapalayam virus. Hantaviruses have a wide geographic distribution and typically cause either hemorrhagic fever with renal syndrome (HFRS) or hantavirus pulmonary syndrome (HPS). Symptoms of HFRS include fever, hemorrhage and renal damage, with a mortality rate up to 15%, depending on the hantavirus species. The first documented case of HFRS occurred in 1934 with a notable epidemic among United Nations soldiers during the Korean War (1951). However, the causative agent of HFRS, Hantaan virus, was not isolated until 1978 (Lee et al. *J. Inf. Dis.*, 1978, 137, 298-308). Symptoms of HPS include fever, pulmonary edema, shock and interstitial pneumonitis (a type of pneumonia involving connective tissue). Sin Nombre virus and Andes virus are two of the hantaviruses that cause a severe form HPS, with an approximately 40% mortality rate. A significant outbreak of pulmonary syndrome occurred in the Southwestern United States in 1993. The etiologic agent of the outbreak was later identified as a hantavirus (Sin Nombre) (Nichol et

al. *Science*, 1993, 262,914-917). The typical route of transmission for hantaviruses is through rodent excreta aerosols, however, Andes virus has been associated with person-to-person transmission ("The Springer Index of Viruses," pgs. 141-174, Tidona and Darai eds., 2001, Springer, New York; Wells et al. *Emerg. Infect. Dis.*, 1997, 3, 171-174).

**[0013]** Members of the phlebovirus genus include Bujaru virus, Chandiru virus, Chilibre virus, Frijoles virus, Punta Toro virus, Rift Valley Fever virus, Salehebad virus, Sandfly fever Naples virus, Uukuniemi virus, Aguacate virus, Anhanga virus, Arboledas virus, Arumowot virus, Caimito virus, Chagres virus, Corfou virus, Gabek Forest virus, Gordil virus, Itaporanga virus, Odrenisrou virus, Pacui virus, Rio Grande virus, Sandfly fever Sicilian virus, Saint-Floris virus and Urucuri virus. Several phleboviruses (e.g., Sandfly fever Naples virus, Sandfly fever Sicilian virus, Chandiru virus and Chagres virus) cause phlebotomus fever, which is typically found in America and the Mediterranean region. Phlebotomus fever, a non-fatal disease, is transmitted by phlebotomines (sand flies) and induces fever, myalgia (muscle pain) and other flu-like symptoms. Rift Valley fever virus, transmitted by mosquitoes, causes a disease of the same name in Africa. Rift Valley fever is characterized by hemorrhagic fever, hepatitis and encephalitis.

**[0014]** Members of the nairovirus genus include Crimean-Congo hemorrhagic fever virus, Dera Ghazi Khan virus, Dugbe virus, Hughes virus, Nairobi sheep disease virus, Qalyub virus, Sakhalin virus and Thiafora virus. Nairoviruses are primarily found in Africa, Asia, Europe and the Middle East. In humans, nairoviruses can cause hemorrhagic fever (Crimean-Congo hemorrhagic fever), Nairobi sheep disease and Dugbe disease. Nairoviruses are typically transmitted to humans by ticks. The first recognized description of Crimean-Congo hemorrhagic fever dates back to the year 1110. This disease is characterized by sudden onset of fever, nausea, severe headache, myalgia and hemorrhage. The mortality rate is approximately 30%. Nairobi sheep disease symptoms include fever, joint pains and general malaise, while Dugbe disease results in fever and prolonged thrombocytopenia (abnormal reduction in platelets) ("The Springer Index of Viruses," pgs. 141-174, Tidona and Darai eds., 2001, Springer, New York).

## **E. Arenaviruses**

**[0015]** Arenavirus is the sole genus of the family Arenaviridae. Arenaviruses are enveloped viruses with a genome that consists of 2 single-stranded RNA segments of negative polarity. The negative-sense RNA of the arenavirus genome serves as both a template for transcription of complementary RNA as well as a template for protein synthesis (ambisense RNA). The genome

segments are designated S, which encodes the nucleocapsid protein (NP) and the precursor glycoprotein (GPC), and L, which encodes the zinc-binding protein (Z) and the RNA-dependent RNA polymerase (L).

[0016] Members of the arenavirus genus include lymphocytic choriomeningitis virus (LCMV), Lassa virus, Ippy virus, Mobala virus, Mopeia virus, Amapari virus, Flexal virus, Guanarito virus, Junin virus, Latino virus, Machupo virus, Parana virus, Pichinde virus, Pirital virus, Oliveros virus, Sabia virus, Tacaribe virus, Tamiami virus, Whitewater Arroyo virus and Pampa virus. A number of arenaviruses are known to cause disease in humans, including LCMV, Lassa virus, Junin virus, Machupo virus, Guanarito virus and Sabia virus. LCMV has a world-wide geographic distribution and infection with LCMV leads to fever, malaise, weakness, myalgia and severe headache. The remaining disease-causing arenaviruses are more limited in their distribution. Lassa fever is found in West Africa and is characterized by fever, headache, dry cough, exudative pharyngitis and hemorrhage. Sabia fever is found in Brazil with symptoms including fever, headache, myalgia (muscle pain), nausea, vomiting and hemorrhage. Junin virus, Machupo virus and Guanarito virus are the causative agents of Argentinean hemorrhagic fever, Bolivian hemorrhagic fever and Venezuelan hemorrhagic fever, respectively, and as their names suggest, are found only in Argentina, Bolivia and Venezuela. Symptoms of these hemorrhagic fevers include malaise, fever, headache, arthralgia (joint pain), nausea, vomiting, hemorrhage and CNS involvement ("The Springer Index of Viruses," pgs. 36-42, Tidona and Darai eds., 2001, Springer, New York).

#### **F. Bioagent Detection**

[0017] A problem in determining the cause of a natural infectious outbreak or a bioterrorist attack is the sheer variety of organisms that can cause human disease. There are over 1400 organisms infectious to humans; many of these have the potential to emerge suddenly in a natural epidemic or to be used in a malicious attack by bioterrorists (Taylor et al., Philos. Trans. R. Soc. London B. Biol. Sci., 2001, 356, 983-989). This number does not include numerous strain variants, bioengineered versions, or pathogens that infect plants or animals.

[0018] Much of the new technology being developed for detection of biological weapons incorporates a polymerase chain reaction (PCR) step based upon the use of highly specific primers and probes designed to selectively detect individual pathogenic organisms. Although this approach is appropriate for the most obvious bioterrorist organisms, like smallpox and anthrax,

experience has shown that it is very difficult to predict which of hundreds of possible pathogenic organisms might be employed in a terrorist attack. Likewise, naturally emerging human disease that has caused devastating consequence in public health has come from unexpected families of bacteria, viruses, fungi, or protozoa. Plants and animals also have their natural burden of infectious disease agents and there are equally important biosafety and security concerns for agriculture.

**[0019]** An alternative to single-agent tests is to do broad-range consensus priming of a gene target conserved across groups of bioagents. Broad-range priming has the potential to generate amplification products across entire genera, families, or, as with bacteria, an entire domain of life. This strategy has been successfully employed using consensus 16S ribosomal RNA primers for determining bacterial diversity, both in environmental samples (Schmidt et al., *J. Bact.*, 1991, 173, 4371-4378) and in natural human flora (Kroes et al., *Proc Nat Acad Sci (USA)*, 1999, 96, 14547-14552). The drawback of this approach for unknown bioagent detection and epidemiology is that analysis of the PCR products requires the cloning and sequencing of hundreds to thousands of colonies per sample, which is impractical to perform rapidly or on a large number of samples.

**[0020]** Conservation of sequence is not as universal for viruses, however, large groups of viral species share conserved protein-coding regions, such as regions encoding viral polymerases or helicases. Like bacteria, consensus priming has also been described for detection of several viral families, including coronaviruses (Stephensen et al., *Vir. Res.*, 1999, 60, 181-189), enteroviruses (Oberste et al., *J. Virol.*, 2002, 76, 1244-51); Oberste et al., *J. Clin. Virol.*, 2003, 26, 375-7); Oberste et al., *Virus Res.*, 2003, 91, 241-8), reovirus (Mack et al., *Proc. Natl. Acad. Sci. U. S. A.*, 1988, 85, 6977-81); Seifarth et al., *AIDS Res. Hum. Retroviruses*, 2000, 16, 721-729); Donehower et al., *J. Vir. Methods*, 1990, 28, 33-46), and adenoviruses (Echavarria et al., *J. Clin. Micro.*, 1998, 36, 3323-3326). However, as with bacteria, there is no adequate analytical method other than sequencing to identify the viral bioagent present.

**[0021]** In contrast to PCR-based methods, mass spectrometry provides detailed information about the molecules being analyzed, including high mass accuracy. It is also a process that can be easily automated. DNA chips with specific probes can only determine the presence or absence of specifically anticipated organisms. Because there are hundreds of thousands of species of

- 8 -

benign pathogens, some very similar in sequence to threat organisms, even arrays with 10,000 probes lack the breadth needed to identify a particular organism.

[0022] There is a need for a method for identification of bioagents which is both specific and rapid, and in which no culture or nucleic acid sequencing is required. Disclosed in U.S. Patent Application Publication Nos. 2003-0027135, 2003-0082539, 2003-0228571, 2004-0209260, 2004-0219517 and 2004-0180328, and in U.S. Application Serial Nos. 10/660,997, 10/728,486, 10/754,415 and 10/829,826,

are methods for identification of bioagents (any organism, cell, or virus, living or dead, or a nucleic acid derived from such an organism, cell or virus) in an unbiased manner by molecular mass and base composition analysis of "bioagent identifying amplicons" which are obtained by amplification of segments of essential and conserved genes which are involved in, for example, translation, replication, recombination and repair, transcription, nucleotide metabolism, amino acid metabolism, lipid metabolism, energy generation, uptake, secretion and the like. Examples of these proteins include, but are not limited to, ribosomal RNAs, ribosomal proteins, DNA and RNA polymerases, RNA-dependent RNA polymerases, RNA capping and methylation enzymes, elongation factors, tRNA synthetases, protein chain initiation factors, heat shock protein groEL, phosphoglycerate kinase, NADH dehydrogenase, DNA ligases, DNA gyrases and DNA topoisomerases, helicases, metabolic enzymes, and the like.

[0023] To obtain bioagent identifying amplicons, primers are selected to hybridize to conserved sequence regions which bracket variable sequence regions to yield a segment of nucleic acid which can be amplified and which is amenable to methods of molecular mass analysis. The variable sequence regions provide the variability of molecular mass which is used for bioagent identification. Upon amplification by PCR or other amplification methods with the specifically chosen primers, an amplification product that represents a bioagent identifying amplicon is obtained. The molecular mass of the amplification product, obtained by mass spectrometry for example, provides the means to uniquely identify the bioagent without a requirement for prior knowledge of the possible identity of the bioagent. The molecular mass of the amplification product or the corresponding base composition (which can be calculated from the molecular mass of the amplification product) is compared with a database of molecular masses or base compositions and a match indicates the identity of the bioagent. Furthermore, the method can be applied to rapid parallel analyses (for example, in a multi-well plate format) the

results of which can be employed in a triangulation identification strategy which is amenable to rapid throughput and does not require nucleic acid sequencing of the amplified target sequence for bioagent identification.

[0024] The result of determination of a previously unknown base composition of a previously unknown bioagent (for example, a newly evolved and heretofore unobserved virus) has downstream utility by providing new bioagent indexing information with which to populate base composition databases. The process of subsequent bioagent identification analyses is thus greatly improved as more base composition data for bioagent identifying amplicons becomes available.

[0025] The present invention provides, *inter alia*, methods of identifying unknown viruses, including viruses of the Filoviridae, Flaviviridae, Bunyaviridae and Arenaviridae families. Also provided are oligonucleotide primers, compositions and kits containing the oligonucleotide primers, which define viral bioagent identifying amplicons and, upon amplification, produce corresponding amplification products whose molecular masses provide the means to identify viruses of the Filoviridae, Flaviviridae, Bunyaviridae and Arenaviridae families at the sub-species level.

## SUMMARY OF THE INVENTION

[0026] The present invention provides primers and compositions comprising pairs of primers, and kits containing the same, and methods for use in identification of viruses in the Filoviridae, Flaviviridae, Bunyaviridae and Arenaviridae families. The primers are designed to produce viral bioagent identifying amplicons of DNA encoding genes essential to virus replication. The invention further provides compositions comprising pairs of primers and kits containing the same, which are designed to provide species and sub-species characterization of members of the Filoviridae, Flaviviridae, Bunyaviridae and Arenaviridae families.

[0027] In some embodiments, an oligonucleotide primer 23 to 35 nucleobases in length comprising at least 70% sequence identity with SEQ ID NO: 129, or a composition comprising the same is provided. In other embodiments, an oligonucleotide primer 22 to 35 nucleobases in length comprising at least 70% sequence identity with SEQ ID NO: 164 is provided. In some embodiments, a composition comprising both primers is provided. In some embodiments, either or both of the primers comprises at least one modified nucleobase, such as a 5-propynyluracil or 5-propynylcytosine. In some embodiments, either or both of the primers comprises at least one

universal nucleobase, such as inosine. In some embodiments, either or both of the primers further comprises a non-templated T residue on the 5'-end. In some embodiments, either or both of the primers comprises at least one non-template tag. In some embodiments, either or both of the primers comprises at least one molecular mass modifying tag. In some embodiments, the foregoing composition(s) are present within a kit. The kit may also comprise at least one calibration polynucleotide, and/or at least one ion exchange resin linked to magnetic beads.

**[0028]** In some embodiments, methods for identification of an unknown filovirus are provided. In some embodiments, nucleic acid from the filovirus is amplified using the composition described above to obtain an amplification product. The molecular mass of the amplification product is measured. Optionally, the base composition of the amplification product is determined from the molecular mass. The molecular mass or base composition is compared with a plurality of molecular masses or base compositions of known filoviral bioagent identifying amplicons, wherein a match between the molecular mass or base composition and a member of the plurality of molecular masses or base compositions identifies the unknown filovirus. In some embodiments, the molecular mass is measured by mass spectrometry.

**[0029]** In some embodiments, methods of determining the presence or absence of a filovirus in a sample are provided. Nucleic acid from the sample is amplified using the composition described above to obtain an amplification product. The molecular mass of the amplification product is determined. Optionally, the base composition of the amplification product is determined from the molecular mass. The molecular mass or base composition of the amplification product is compared with the known molecular masses or base compositions of one or more known filoviral bioagent identifying amplicons, wherein a match between the molecular mass or base composition of the amplification product and the molecular mass or base composition of one or more known filoviral bioagent identifying amplicons indicates the presence of the filovirus in the sample. In some embodiments, the molecular mass is measured by mass spectrometry.

**[0030]** In some embodiments, methods for determination of the quantity of an unknown filovirus in a sample are provided. The sample is contacted with the composition described above and a known quantity of a calibration polynucleotide comprising a calibration sequence. Nucleic acid from the unknown filovirus in the sample is concurrently amplified with the composition described above and nucleic acid from the calibration polynucleotide in the sample

is concurrently amplified with the composition described above to obtain a first amplification product comprising a filoviral bioagent identifying amplicon and a second amplification product comprising a calibration amplicon. The molecular mass and abundance for the filoviral bioagent identifying amplicon and the calibration amplicon is determined. The filoviral bioagent identifying amplicon is distinguished from the calibration amplicon based on molecular mass, wherein comparison of filoviral bioagent identifying amplicon abundance and calibration amplicon abundance indicates the quantity of filovirus in the sample. In some embodiments, the base composition of the filoviral bioagent identifying amplicon is determined.

### BRIEF DESCRIPTION OF THE DRAWINGS

[0031] The foregoing summary of the invention, as well as the following detailed description of the invention, is better understood when read in conjunction with the accompanying drawings which are included by way of example and not by way of limitation.

[0032] **Figure 1** is a process diagram illustrating a representative primer selection process.

[0033] **Figure 2** is a graph of the inverse figure of merit  $\phi$  plotted for a master list of 16 primer sets in a *Yersinia pestis* target biocluster.

[0034] **Figure 3** is a graph showing the base compositions of the 229E Human Coronavirus, OC43 Human Coronavirus and the SARS Coronavirus.

[0035] **Figure 4** shows the phylogenetic relationship between a number of animal coronavirus species.

[0036] **Figure 5A** is a flow chart illustrating a method of training an embodiment of a polytope pattern classifier; **Figure 5B** is a flow chart illustrating a method of identifying an unknown sample using an embodiment of a trained polytope pattern classifier.

[0037] **Figure 6A** is a flow chart illustrating a method of training an embodiment of a polytope pattern classifier of a lower dimension when the sample space is reduced in dimension by imposing a constraint. **Figure 6B** and **Figure 6C** are flow charts illustrating the method of identifying a unknown bioagent using different embodiments of a trained polytope pattern classifier.

[0038] **Figure 7A** is a three dimensional representation of a polytope defined by applying the three unary inequality constraints; **Figure 7B** and **Figure 7C** are three dimensional representations of polytopes defined by additionally applying a unary inequality on A, equivalent to a trinary inequality on the three dimensions shown.

[0039] **Figure 8A** and **Figure 8B** are three dimensional representations of polytopes defined by applying the C+T (pyrimidine/purine) binary inequality.

[0040] **Figure 9A** and **Figure 9 B** are three dimensional representations of polytopes defined by applying the G+T (keto/amino preference) binary inequality.

[0041] **Figure 10** is a three dimensional representation of polytopes defined by applying the G+C (strong/weak base paring constraints).

[0042] **Figure 11A** shows the three dimensional representation of the *Neisseriales* polytope along with its population, volume and density; **Figure 11B** shows the addition of the three dimensional representation of the Nitrosomonades polytope along with its population, volume and density to the polytope of **Figure 11A**; **Figure 11C** shows the addition of the three dimensional representation of the Burkholderiales polytope along with its population, volume and density to the polytope of **Figure 11B**; **Figure 11D** shows the addition of the three dimensional representation of the Hydrogenophilales polytope along with its population, volume and density; to the polytope of **Figure 11C**; **Figure 11E** shows the addition of the three dimensional representation of the Rhodocyclales polytope along with its population, volume and density to the polytope of **Figure 11D**; **Figure 11F** outlines the polytope for betaproteobacteria order in relationship to the five exemplary taxons.

[0043] **Figure 12** is a comparison of the individual probabilities of detecting a bioagent using individual amplicons as compared to the overall probability of classifying the bioagent using multiple amplicons.

[0044] **Figure 13** is an graph illustrating the reliability of phylogenetic assignment made using one embodiment of the polytope pattern classifier.

[0045] **Figure 14** is a process diagram illustrating an embodiment of the calibration method.

**DETAILED DESCRIPTION OF EMBODIMENTS**

[0046] In the context of the present invention, a “bioagent” is any organism, cell, or virus, living or dead, or a nucleic acid derived from such an organism, cell or virus. Examples of bioagents include, but are not limited, to cells, including but not limited to human clinical samples, cell cultures, bacterial cells and other pathogens), viruses, viroids, fungi, protists, parasites, and pathogenicity markers (including, but not limited to: pathogenicity islands, antibiotic resistance genes, virulence factors, toxin genes and other bioregulating compounds). Samples may be alive or dead or in a vegetative state (for example, vegetative bacteria or spores) and may be encapsulated or bioengineered. In the context of this invention, a “pathogen” is a bioagent which causes a disease or disorder.

[0047] As used herein, “intelligent primers” are primers that are designed to bind to highly conserved sequence regions of a bioagent identifying amplicon that flank an intervening variable region and yield amplification products which ideally provide enough variability to distinguish each individual bioagent, and which are amenable to molecular mass analysis. By the term “highly conserved,” it is meant that the sequence regions exhibit between about 80-100%, or between about 90-100%, or between about 95-100% identity among all or at least 70%, at least 80%, at least 90%, at least 95%, or at least 99% of species or strains.

[0048] As used herein, “broad range survey primers” are intelligent primers designed to identify an unknown bioagent as a member of a particular division (e.g., an order, family, class, clade, genus or other such grouping of bioagents above the species level of bioagents). In some cases, broad range survey primers are able to identify unknown bioagents at the species or sub-species level. As used herein, “division-wide primers” are intelligent primers designed to identify a bioagent at the species level and “drill-down” primers are intelligent primers designed to identify a bioagent at the sub-species level. As used herein, the “sub-species” level of identification includes, but is not limited to, strains, subtypes, variants, and isolates.

[0049] As used herein, a “bioagent division” is defined as group of bioagents above the species level and includes but is not limited to, orders, families, classes, clades, genera or other such groupings of bioagents above the species level.

[0050] As used herein, a “sub-species characteristic” is a genetic characteristic that provides the means to distinguish two members of the same bioagent species. For example, one viral

- 14 -

strain could be distinguished from another viral strain of the same species by possessing a genetic change (e.g., for example, a nucleotide deletion, addition or substitution) in one of the viral genes, such as the RNA-dependent RNA polymerase. In this case, the sub-species characteristic that can be identified using the methods of the present invention, is the genetic change in the viral polymerase.

[0051] As used herein, the term “bioagent identifying amplicon” refers to a polynucleotide that is amplified from a bioagent in an amplification reaction and which 1) provides enough variability to distinguish each individual bioagent and 2) whose molecular mass is amenable to molecular mass determination.

[0052] As used herein, a “base composition” is the exact number of each nucleobase (A, T, C and G) in a given sequence.

[0053] As used herein, a “base composition signature” (BCS) is the exact base composition (i.e., the number of A, T, G and C nucleobases) determined from the molecular mass of a bioagent identifying amplicon.

[0054] As used herein, a “base composition probability cloud” is a representation of the diversity in base composition resulting from a variation in sequence that occurs among different isolates of a given species. The “base composition probability cloud” represents the base composition constraints for each species and is typically visualized using a pseudo four-dimensional plot.

[0055] As used herein, a “wobble base” is a variation in a codon found at the third nucleotide position of a DNA triplet. Variations in conserved regions of sequence are often found at the third nucleotide position due to redundancy in the amino acid code.

[0056] In the context of the present invention, the term “unknown bioagent” may mean either: (i) a bioagent whose existence is known (such as the well known bacterial species *Staphylococcus aureus* for example) but which is not known to be in a sample to be analyzed, or (ii) a bioagent whose existence is not known (for example, the SARS coronavirus was unknown prior to April 2003). For example, if the method for identification of coronaviruses disclosed in commonly owned U.S. Patent Serial No. 10/829,826

- 15 -

was to be employed prior to April 2003 to identify the SARS coronavirus in a clinical sample, both meanings of “unknown” bioagent are applicable since the SARS coronavirus was unknown to science prior to April, 2003 and since it was not known what bioagent (in this case a coronavirus) was present in the sample. On the other hand, if the method of U.S. Patent Serial No. 10/829,826 was to be employed subsequent to April 2003 to identify the SARS coronavirus in a clinical sample, only the first meaning (i) of “unknown” bioagent would apply since the SARS coronavirus became known to science subsequent to April 2003 and since it was not known what bioagent was present in the sample.

[0057] As used herein, “triangulation identification” means the employment of more than one bioagent identifying amplicons for identification of a bioagent.

[0058] In the context of the present invention, “viral nucleic acid” includes, but is not limited to, DNA, RNA, or DNA that has been obtained from viral RNA, such as, for example, by performing a reverse transcription reaction. Viral RNA can either be single-stranded (of positive or negative polarity) or double-stranded.

[0059] As used herein, the term “etiology” refers to the causes or origins, of diseases or abnormal physiological conditions.

[0060] As used herein, the term “nucleobase” is synonymous with other terms in use in the art including “nucleotide,” “deoxynucleotide,” “nucleotide residue,” “deoxynucleotide residue,” “nucleotide triphosphate (NTP),” or deoxynucleotide triphosphate (dNTP).

[0061] The present invention provides methods for detection and identification of bioagents in an unbiased manner using bioagent identifying amplicons. Intelligent primers are selected to hybridize to conserved sequence regions of nucleic acids derived from a bioagent and which bracket variable sequence regions to yield a bioagent identifying amplicon which can be amplified and which is amenable to molecular mass determination. The molecular mass then provides a means to uniquely identify the bioagent without a requirement for prior knowledge of the possible identity of the bioagent. The molecular mass or corresponding base composition signature (BCS) of the amplification product is then matched against a database of molecular masses or base composition signatures. Furthermore, the method can be applied to rapid parallel multiplex analyses, the results of which can be employed in a triangulation identification

strategy. The present method provides rapid throughput and does not require nucleic acid sequencing of the amplified target sequence for bioagent detection and identification.

[0062] Despite enormous biological diversity, all forms of life on earth share sets of essential, common features in their genomes. Since genetic data provide the underlying basis for identification of bioagents by the methods of the present invention, it is necessary to select segments of nucleic acids which ideally provide enough variability to distinguish each individual bioagent and whose molecular mass is amenable to molecular mass determination.

[0063] Unlike bacterial genomes, which exhibit conservation of numerous genes (i.e. housekeeping genes) across all organisms, viruses do not share a gene that is essential and conserved among all virus families. Therefore, viral identification is achieved within smaller groups of related viruses, such as members of a particular virus family or genus. For example, RNA-dependent RNA polymerase is present in all single-stranded RNA viruses and can be used for broad priming as well as resolution within the virus family.

[0064] In some embodiments of the present invention, at least one viral nucleic acid segment is amplified in the process of identifying the bioagent. Thus, the nucleic acid segments that can be amplified by the primers disclosed herein and that provide enough variability to distinguish each individual bioagent and whose molecular masses are amenable to molecular mass determination are herein described as bioagent identifying amplicons.

[0065] In some embodiments of the present invention, bioagent identifying amplicons comprise from about 45 to about 200 nucleobases (i.e. from about 45 to about 200 linked nucleosides). One of ordinary skill in the art will appreciate that the invention embodies compounds of 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, and 200 nucleobases in length, or any range therewithin.

[0066] It is the combination of the portions of the bioagent nucleic acid segment to which the primers hybridize (hybridization sites) and the variable region between the primer hybridization sites that comprises the bioagent identifying amplicon. In some embodiments, bioagent identifying amplicons amenable to molecular mass determination which are produced by the primers described herein are either of a length, size or mass compatible with the particular mode of molecular mass determination or compatible with a means of providing a predictable fragmentation pattern in order to obtain predictable fragments of a length compatible with the particular mode of molecular mass determination. Such means of providing a predictable fragmentation pattern of an amplification product include, but are not limited to, cleavage with restriction enzymes or cleavage primers, for example. Thus, in some embodiments, bioagent identifying amplicons are larger than 200 nucleobases and are amenable to molecular mass determination following restriction digestion. Methods of using restriction enzymes and cleavage primers are well known to those with ordinary skill in the art.

[0067] In some embodiments, amplification products corresponding to bioagent identifying amplicons are obtained using the polymerase chain reaction (PCR) which is a routine method to those with ordinary skill in the molecular biology arts. Other amplification methods may be used such as ligase chain reaction (LCR), low-stringency single primer PCR, and multiple strand displacement amplification (MDA) which are also well known to those with ordinary skill.

[0068] Intelligent primers are designed to bind to highly conserved sequence regions of a bioagent identifying amplicon that flank an intervening variable region and yield amplification products which ideally provide enough variability to distinguish each individual bioagent, and which are amenable to molecular mass analysis. In some embodiments, the highly conserved sequence regions exhibit between about 80-100%, or between about 90-100%, or between about 95-100% identity, or between about 99-100% identity. The molecular mass of a given amplification product provides a means of identifying the bioagent from which it was obtained, due to the variability of the variable region. Thus design of intelligent primers requires selection of a variable region with appropriate variability to resolve the identity of a given bioagent. Bioagent identifying amplicons are ideally specific to the identity of the bioagent.

[0069] Identification of bioagents can be accomplished at different levels using intelligent primers suited to resolution of each individual level of identification. Broad range survey intelligent primers are designed with the objective of identifying a bioagent as a member of a

particular division (e.g., an order, family, class, clade, genus or other such grouping of bioagents above the species level of bioagents). As a non-limiting example, members of the filovirus genus may be identified as such by employing broad range survey intelligent primers such as primers which target the viral RNA-dependent RNA polymerase. As another non-limiting example, members of the hantavirus genus may be identified as such by employing broad range survey intelligent primers such as primers which target the viral RNA-dependent RNA polymerase. In some embodiments, broad range survey intelligent primers are capable of identification of bioagents at the species or sub-species level.

**[0070]** Division-wide intelligent primers are designed with an objective of identifying a bioagent at the species level. As a non-limiting example, Zaire Ebola virus, Sudan Ebola virus and Marburg virus, species of the filovirus genus, can be distinguished from each other using division-wide intelligent primers. As another non-limiting example, Hantaan, Sin Nombre and Andes virus, species of the hantavirus genus, can be distinguished from each other using division-wide intelligent primers. Division-wide intelligent primers are not always required for identification at the species level because broad range survey intelligent primers may provide sufficient identification resolution to accomplishing this identification objective.

**[0071]** Drill-down intelligent primers are designed with the objective of identifying a bioagent at the sub-species level (including strains, subtypes, variants and isolates) based on sub-species characteristics. As one non-limiting example, the Mayinga, Zaire and Eckron isolates of Zaire Ebola can be distinguished from each other using drill-down primers. As another non-limiting example, the NMR11, NMH10 and CC107 isolates of Sin Nombre virus can be distinguished from each other using drill-down primers. Drill-down intelligent primers are not always required for identification at the sub-species level because broad range survey intelligent primers may provide sufficient identification resolution to accomplishing this identification objective.

**[0072]** A representative process flow diagram used for primer selection and validation process is outlined in Figure 1. For each group of organisms, candidate target sequences are identified (200) from which nucleotide alignments are created (210) and analyzed (220). Primers are then designed by selecting appropriate priming regions (230) which then makes possible the selection of candidate primer pairs (240). The primer pairs are then subjected to *in silico* analysis by electronic PCR (ePCR) (300) wherein bioagent identifying amplicons are obtained from sequence databases such as GenBank or other sequence collections (310) and checked for

- 19 -

specificity *in silico* (340). Bioagent identifying amplicons obtained from GenBank sequences (310) can also be analyzed by a probability model (320) which predicts the capability of a given amplicon to identify unknown bioagents such that the base compositions of amplicons with favorable probability scores are then stored in a base composition database (325). Alternatively, base compositions of the bioagent identifying amplicons obtained from the primers and GenBank sequences can be directly entered into the base composition database (330). Candidate primer pairs (240) are validated by *in vitro* amplification by a method such as PCR analysis (400) of nucleic acid from a collection of organisms (410). Amplification products thus obtained are analyzed to confirm the sensitivity, specificity and reproducibility of the primers used to obtain the amplification products (420).

[0073] Many of the important pathogens, including the organisms of greatest concern as biological weapons agents, have been completely sequenced. This effort has greatly facilitated the design of primers and probes for the detection of unknown bioagents. The combination of broad-range priming with division-wide and drill-down priming has been used very successfully in several applications of the technology, including environmental surveillance for biowarfare threat agents and clinical sample analysis for medically important pathogens.

[0074] Synthesis of primers is well known and routine in the art. The primers may be conveniently and routinely made through the well-known technique of solid phase synthesis. Equipment for such synthesis is sold by several vendors including, for example, Applied Biosystems (Foster City, CA). Any other means for such synthesis known in the art may additionally or alternatively be employed.

[0075] The primers are employed as compositions for use in methods for identification of viral bioagents as follows: a primer pair composition is contacted with nucleic acid (such as, for example, DNA from a DNA virus, or DNA reverse transcribed from the RNA of an RNA virus) of an unknown viral bioagent. The nucleic acid is then amplified by a nucleic acid amplification technique, such as PCR for example, to obtain an amplification product that represents a bioagent identifying amplicon. The molecular mass of each strand of the double-stranded amplification product is determined by a molecular mass measurement technique such as mass spectrometry for example, wherein the two strands of the double-stranded amplification product are separated during the ionization process. In some embodiments, the mass spectrometry is electrospray Fourier transform ion cyclotron resonance mass spectrometry (ESI-FTICR-MS) or

electrospray time of flight mass spectrometry (ESI-TOF-MS). A list of possible base compositions can be generated for the molecular mass value obtained for each strand and the choice of the correct base composition from the list is facilitated by matching the base composition of one strand with a complementary base composition of the other strand. The molecular mass or base composition thus determined is then compared with a database of molecular masses or base compositions of analogous bioagent identifying amplicons for known viral bioagents. A match between the molecular mass or base composition of the amplification product and the molecular mass or base composition of an analogous bioagent identifying amplicon for a known viral bioagent indicates the identity of the unknown bioagent. In some embodiments, the primer pair used is one of the primer pairs of Tables 4-7. In some embodiments, the method is repeated using a different primer pair to resolve possible ambiguities in the identification process or to improve the confidence level for the identification assignment.

[0076] In some embodiments, a bioagent identifying amplicon may be produced using only a single primer (either the forward or reverse primer of any given primer pair), provided an appropriate amplification method is chosen, such as, for example, low stringency single primer PCR (LSSP-PCR). Adaptation of this amplification method in order to produce bioagent identifying amplicons can be accomplished by one with ordinary skill in the art without undue experimentation.

[0077] In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding the RNA-dependent RNA polymerase of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known filoviruses and produce bioagent identifying amplicons. In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding nucleocapsid of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known filoviruses and produce bioagent identifying amplicons.

[0078] In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding the RNA-dependent RNA polymerase (NS5) of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known flaviviruses and produce bioagent identifying

amplicons. In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding the protease/helicase (NS3) of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known flaviviruses and produce bioagent identifying amplicons.

[0079] In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding the RNA-dependent RNA polymerase of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known hantaviruses and produce bioagent identifying amplicons. In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding nucleocapsid of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known hantaviruses and produce bioagent identifying amplicons.

[0080] In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding the RNA-dependent RNA polymerase of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known phleboviruses and produce bioagent identifying amplicons.

[0081] In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding nucleocapsid of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known nairoviruses and produce bioagent identifying amplicons.

[0082] In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding the RNA-dependent RNA polymerase (L) of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known arenaviruses and produce bioagent identifying amplicons. In some embodiments, the oligonucleotide primers are broad range survey primers which hybridize to conserved regions of nucleic acid encoding nucleocapsid (NP) of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known arenaviruses and produce bioagent identifying amplicons.

**[0083]** As used herein, the term broad range survey primers refers to primers that bind to nucleic acid encoding genes essential to filovirus, flavivirus, hantavirus, phlebovirus, nairovirus or arenavirus replication (e.g., for example, RNA-dependent RNA polymerase or nucleocapsid) of all (or between 80% and 100%, between 85% and 100%, between 90% and 100% or between 95% and 100%) known species of filovirus, flavivirus, hantavirus, phlebovirus, nairovirus or arenavirus.

**[0084]** In some embodiments, the broad range survey primer pairs comprise oligonucleotides ranging in length from 13-35 nucleobases, each of which have from 70% to 100% sequence identity with primer pair number 853, which corresponds to SEQ ID NOs: 129:164. In some embodiments, the broad range survey primer pairs comprise oligonucleotides ranging in length from 13-35 nucleobases, each of which have from 70% to 100% sequence identity with primer pair number 858, which corresponds to SEQ ID NOs: 124:159. In some embodiments, the broad range survey primer pairs comprise oligonucleotides ranging in length from 13-35 nucleobases, each of which have from 70% to 100% sequence identity with primer pair number 856, which corresponds to SEQ ID NOs: 134:169. In some embodiments, the broad range survey primer pairs comprise oligonucleotides ranging in length from 13-35 nucleobases, each of which have from 70% to 100% sequence identity with primer pair number 864, which corresponds to SEQ ID NOs: 138:174.

**[0085]** In some cases, the molecular mass or base composition of a viral bioagent identifying amplicon defined by a broad range survey primer pair does not provide enough resolution to unambiguously identify a viral bioagent at the species level. These cases benefit from further analysis of one or more viral bioagent identifying amplicons generated from at least one additional broad range survey primer pair or from at least one additional division-wide primer pair. The employment of more than one bioagent identifying amplicon for identification of a bioagent is herein referred to as triangulation identification.

**[0086]** In other embodiments, the oligonucleotide primers are division-wide primers which hybridize to nucleic acid encoding genes of species within a genus of viruses. In other embodiments, the oligonucleotide primers are drill-down primers which enable the identification of sub-species characteristics. Drill down primers provide the functionality of producing bioagent identifying amplicons for drill-down analyses such as strain typing when contacted with nucleic acid under amplification conditions. Identification of such sub-species characteristics is

often critical for determining proper clinical treatment of viral infections. In some embodiments, sub-species characteristics are identified using only broad range survey primers and division-wide and drill-down primers are not used.

[0087] In some embodiments, the primers used for amplification hybridize to and amplify genomic DNA, DNA of bacterial plasmids, DNA of DNA viruses or DNA reverse transcribed from RNA of an RNA virus.

[0088] In some embodiments, the primers used for amplification hybridize directly to viral RNA and act as reverse transcription primers for obtaining DNA from direct amplification of viral RNA. Methods of amplifying RNA using reverse transcriptase are well known to those with ordinary skill in the art and can be routinely established without undue experimentation.

[0089] One with ordinary skill in the art of design of amplification primers will recognize that a given primer need not hybridize with 100% complementarity in order to effectively prime the synthesis of a complementary nucleic acid strand in an amplification reaction. Moreover, a primer may hybridize over one or more segments such that intervening or adjacent segments are not involved in the hybridization event. (e.g., for example, a loop structure or a hairpin structure). The primers of the present invention may comprise at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95% or at least 99% sequence identity with any of the primers listed in Tables 4-7. Thus, in some embodiments of the present invention, an extent of variation of 70% to 100%, or any range therewithin, of the sequence identity is possible relative to the specific primer sequences disclosed herein. Determination of sequence identity is described in the following example: a primer 20 nucleobases in length which is identical to another 20 nucleobase primer having two non-identical residues has 18 of 20 identical residues ( $18/20 = 0.9$  or 90% sequence identity). In another example, a primer 15 nucleobases in length having all residues identical to a 15 nucleobase segment of primer 20 nucleobases in length would have  $15/20 = 0.75$  or 75% sequence identity with the 20 nucleobase primer.

[0090] Percent homology, sequence identity or complementarity, can be determined by, for example, the Gap program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, Madison WI), using default settings, which uses the algorithm of Smith and Waterman (Adv. Appl. Math., 1981, 2, 482-489). In some embodiments, complementarity of primers with respect to the conserved priming regions of viral nucleic acid,

is between about 70% and about 80%. In other embodiments, homology, sequence identity or complementarity, is between about 80% and about 90%. In yet other embodiments, homology, sequence identity or complementarity, is at least 90%, at least 92%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99% or is 100%.

[0091] In some embodiments, the primers described herein comprise at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 92%, at least 94%, at least 95%, at least 96%, at least 98%, or at least 99%, or 100% (or any range therewithin) sequence identity with the primer sequences specifically disclosed herein. Thus, for example, a primer may have between 70% and 100%, between 75% and 100%, between 80% and 100%, and between 95% and 100% sequence identity with SEQ ID NO: 129. Likewise, a primer may have similar sequence identity with any other primer whose nucleotide sequence is disclosed herein.

[0092] One with ordinary skill is able to calculate percent sequence identity or percent sequence homology and able to determine, without undue experimentation, the effects of variation of primer sequence identity on the function of the primer in its role in priming synthesis of a complementary strand of nucleic acid for production of an amplification product of a corresponding bioagent identifying amplicon.

[0093] In some embodiments of the present invention, the oligonucleotide primers are 13 to 35 nucleobases in length (13 to 35 linked nucleotide residues). These embodiments comprise oligonucleotide primers 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 or 35 nucleobases in length, or any range therewithin.

[0094] In some embodiments, any given primer comprises a modification comprising the addition of a non-templated T residue to the 5' end of the primer (i.e., the added T residue does not necessarily hybridize to the nucleic acid being amplified). The addition of a non-templated T residue has an effect of minimizing the addition of non-templated A residues as a result of the non-specific enzyme activity of *Taq* polymerase (Magnuson et al., *Biotechniques*, 1996, 21, 700-709), an occurrence which may lead to ambiguous results arising from molecular mass analysis.

[0095] In some embodiments of the present invention, primers may contain one or more universal bases. Because any variation (due to codon wobble in the 3<sup>rd</sup> position) in the conserved regions among species is likely to occur in the third position of a DNA (or RNA) triplet,

- 25 -

oligonucleotide primers can be designed such that the nucleotide corresponding to this position is a base which can bind to more than one nucleotide, referred to herein as a "universal nucleobase." For example, under this "wobble" pairing, inosine (I) binds to U, C or A; guanine (G) binds to U or C, and uridine (U) binds to U or C. Other examples of universal nucleobases include nitroindoles such as 5-nitroindole or 3-nitropyrrole (Loakes et al., *Nucleosides and Nucleotides*, 1995, 14, 1001-1003), the degenerate nucleotides dP or dK (Hill *et al.*), an acyclic nucleoside analog containing 5-nitroindazole (Van Aerschot et al., *Nucleosides and Nucleotides*, 1995, 14, 1053-1056) or the purine analog 1-(2-deoxy- $\beta$ -D-ribofuranosyl)-imidazole-4-carboxamide (Sala et al., *Nucl. Acids Res.*, 1996, 24, 3302-3306).

[0096] In some embodiments, to compensate for the somewhat weaker binding by the wobble base, the oligonucleotide primers are designed such that the first and second positions of each triplet are occupied by nucleotide analogs which bind with greater affinity than the unmodified nucleotide. Examples of these analogs include, but are not limited to, 2,6-diaminopurine which binds to thymine, 5-propynyluracil which binds to adenine and 5-propynylcytosine and phenoxazines, including G-clamp, which binds to G. Propynylated pyrimidines are described in U.S. Patent Nos. 5,645,985, 5,830,653 and 5,484,908, each of which is commonly owned and incorporated herein by reference in its entirety. Propynylated primers are described in U.S. Pre-Grant Publication No. 2003-0170682, which is also commonly owned and incorporated herein by reference in its entirety. Phenoxazines are described in U.S. Patent Nos. 5,502,177, 5,763,588, and 6,005,096. G-clamps are described in U.S. Patent Nos. 6,007,992 and 6,028,183.

[0097] In some embodiments, to enable broad priming of rapidly evolving RNA viruses, primer hybridization is enhanced using primers and probes containing 5-propynyl deoxy-cytidine and deoxy-thymidine nucleotides. These modified primers and probes offer increased affinity and base pairing selectivity.

[0098] In some embodiments, non-template primer tags are used to increase the melting temperature ( $T_m$ ) of a primer-template duplex in order to improve amplification efficiency. A non-template tag is at least three consecutive A or T nucleotide residues on a primer which are not complementary to the template. In any given non-template tag, A can be replaced by C or G and T can also be replaced by C or G. Although Watson-Crick hybridization is not expected to

occur for a non-template tag relative to the template, the extra hydrogen bond in a G-C pair relative to an A-T pair confers increased stability of the primer-template duplex and improves amplification efficiency for subsequent cycles of amplification when the primers hybridize to strands synthesized in previous cycles.

**[0099]** In other embodiments, propynylated tags may be used in a manner similar to that of the non-template tag, wherein two or more 5-propynylcytidine or 5-propynyluridine residues replace template matching residues on a primer. In other embodiments, a primer contains a modified internucleoside linkage such as a phosphorothioate linkage, for example.

**[0100]** In some embodiments, the primers contain mass-modifying tags. Reducing the total number of possible base compositions of a nucleic acid of specific molecular weight provides a means of avoiding a persistent source of ambiguity in determination of base composition of amplification products. Addition of mass-modifying tags to certain nucleobases of a given primer will result in simplification of *de novo* determination of base composition of a given bioagent identifying amplicon from its molecular mass.

**[0101]** In some embodiments of the present invention, the mass modified nucleobase comprises one or more of the following: for example, 7-deaza-2'-deoxyadenosine-5'-triphosphate, 5-iodo-2'-deoxyuridine-5'-triphosphate, 5-bromo-2'-deoxyuridine-5'-triphosphate, 5-bromo-2'-deoxycytidine-5'-triphosphate, 5-iodo-2'-deoxycytidine-5'-triphosphate, 5-hydroxy-2'-deoxyuridine-5'-triphosphate, 4-thiothymidine-5'-triphosphate, 5-aza-2'-deoxyuridine-5'-triphosphate, 5-fluoro-2'-deoxyuridine-5'-triphosphate, O6-methyl-2'-deoxyguanosine-5'-triphosphate, N2-methyl-2'-deoxyguanosine-5'-triphosphate, 8-oxo-2'-deoxyguanosine-5'-triphosphate or thiothymidine-5'-triphosphate. In some embodiments, the mass-modified nucleobase comprises  $^{15}\text{N}$  or  $^{13}\text{C}$  or both  $^{15}\text{N}$  and  $^{13}\text{C}$ .

**[0102]** In some cases, a molecular mass of a given bioagent identifying amplicon alone does not provide enough resolution to unambiguously identify a given bioagent. The employment of more than one bioagent identifying amplicon for identification of a bioagent is herein referred to as triangulation identification. Triangulation identification is pursued by analyzing a plurality of bioagent identifying amplicons selected within multiple core genes. This process is used to reduce false negative and false positive signals, and enable reconstruction of the origin of hybrid or otherwise engineered bioagents. For example, identification of the three part toxin genes

typical of *B. anthracis* (Bowen et al., J. Appl. Microbiol., 1999, 87, 270-278) in the absence of the expected signatures from the *B. anthracis* genome would suggest a genetic engineering event.

[0103] In some embodiments, the triangulation identification process can be pursued by characterization of bioagent identifying amplicons in a massively parallel fashion using the polymerase chain reaction (PCR), such as multiplex PCR where multiple primers are employed in the same amplification reaction mixture, or PCR in multi-well plate format wherein a different and unique pair of primers is used in multiple wells containing otherwise identical reaction mixtures. Such multiplex and multi-well PCR methods are well known to those with ordinary skill in the arts of rapid throughput amplification of nucleic acids.

[0104] In some embodiments, the molecular mass of a given bioagent identifying amplicon is determined by mass spectrometry. Mass spectrometry has several advantages, not the least of which is high bandwidth characterized by the ability to separate (and isolate) many molecular peaks across a broad range of mass to charge ratio ( $m/z$ ). Thus mass spectrometry is intrinsically a parallel detection scheme without the need for radioactive or fluorescent labels, since every amplification product is identified by its molecular mass. The current state of the art in mass spectrometry is such that less than femtomole quantities of material can be readily analyzed to afford information about the molecular contents of the sample. An accurate assessment of the molecular mass of the material can be quickly obtained, irrespective of whether the molecular weight of the sample is several hundred, or in excess of one hundred thousand atomic mass units (amu) or Daltons.

[0105] In some embodiments, intact molecular ions are generated from amplification products using one of a variety of ionization techniques to convert the sample to gas phase. These ionization methods include, but are not limited to, electrospray ionization (ESI), matrix-assisted laser desorption ionization (MALDI) and fast atom bombardment (FAB). Upon ionization, several peaks are observed from one sample due to the formation of ions with different charges. Averaging the multiple readings of molecular mass obtained from a single mass spectrum affords an estimate of molecular mass of the bioagent identifying amplicon. Electrospray ionization mass spectrometry (ESI-MS) is particularly useful for very high molecular weight polymers such as proteins and nucleic acids having molecular weights greater than 10 kDa, since it yields a

distribution of multiply-charged molecules of the sample without causing a significant amount of fragmentation.

[0106] The mass detectors used in the methods of the present invention include, but are not limited to, Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS), time of flight (TOF), ion trap, quadrupole, magnetic sector, Q-TOF, and triple quadrupole.

[0107] Although the molecular mass of amplification products obtained using intelligent primers provides a means for identification of bioagents, conversion of molecular mass data to a base composition signature is useful for certain analyses. As used herein, a base composition signature (BCS) is the exact base composition determined from the molecular mass of a bioagent identifying amplicon. In one embodiment, a BCS provides an index of a specific gene in a specific organism.

[0108] In some embodiments, conversion of molecular mass data to a base composition is useful for certain analyses. As used herein, a base composition is the exact number of each nucleobase (A, T, C and G).

[0109] RNA viruses depend on error-prone polymerases for replication and therefore their nucleotide sequences (and resultant base compositions) drift over time within the functional constraints allowed by selection pressure. Base composition probability distribution of a viral species or group represents a probabilistic distribution of the above variation in the A, C, G and T base composition space and can be derived by analyzing base compositions of all known isolates of that particular species.

[0110] In some embodiments, assignment of base compositions to experimentally determined molecular masses is accomplished using base composition probability clouds. Base compositions, like sequences, vary slightly from isolate to isolate within species. It is possible to manage this diversity by building base composition probability clouds around the composition constraints for each species. This permits identification of organisms in a fashion similar to sequence analysis. A pseudo four-dimensional plot can be used to visualize the concept of base composition probability clouds. Optimal primer design requires optimal choice of bioagent identifying amplicons and maximizes the separation between the base composition signatures of individual bioagents. Areas where clouds overlap indicate regions that may result in a

misclassification, a problem which is overcome by a triangulation identification process using bioagent identifying amplicons not affected by overlap of base composition probability clouds.

[0111] In some embodiments, base composition probability clouds provide the means for screening potential primer pairs in order to avoid potential misclassifications of base compositions. In other embodiments, base composition probability clouds provide the means for predicting the identity of a bioagent whose assigned base composition was not previously observed and/or indexed in a bioagent identifying amplicon base composition database due to evolutionary transitions in its nucleic acid sequence. Thus, in contrast to probe-based techniques, mass spectrometry determination of base composition does not require prior knowledge of the composition or sequence in order to make the measurement.

[0112] The present invention provides bioagent classifying information similar to DNA sequencing and phylogenetic analysis at a level sufficient to identify a given bioagent. Furthermore, the process of determination of a previously unknown base composition for a given bioagent (for example, in a case where sequence information is unavailable) has downstream utility by providing additional bioagent indexing information with which to populate base composition databases. The process of future bioagent identification is thus greatly improved as more base composition indexes become available in base composition databases.

[0113] Existing nucleic acid-based tests for bioagent detection are primarily based upon amplification methods using primer and probes designed to detect specific organisms. Because prior knowledge of nucleic acid sequence information is required to develop these probe-based tests they cannot be used to identify unanticipated, newly emergent, or previously unknown infections organisms. Thus, the discovery of new bioagents still relies largely on traditional culture methods and microscopy.

[0114] Methods of the present invention, however, allow rapid identification of new bioagent species without the need for prior knowledge of nucleotide sequence. This is achieved by applying a mathematical and/or probabilistic model for sequence variation developed based on known bioagent amplicon base composition (the “training set” of data) and matching the unknown bioagent data (“test data”) to the model.

[0115] For unambiguous detection and identification of bioagents, it would be ideal if every isolate of a given species of bioagent (*E. coli*, for example) had exactly the same base count in any particular amplified region. However, due to naturally occurring mutations and/or deliberately engineered changes, isolates of any species might have some variation in the base count of a particular region. Because of naturally occurring variation and because engineered threat bioagents may differ slightly in particular regions from their naturally occurring counterparts, it is useful to “blur” the expected base count for a given species to allow for this variation so that the system does not miss detections. The more the expected base count is blurred, the less likely it is that a particular species will escape detection; however, such blurring will cause more overlap between the expected base counts of different species, contributing to misclassifications.

[0116] To solve this problem, expected base counts can be blurred according to the natural principles of biological mutations, customizing the specific blurring to the biological constraints of each amplified region. Each amplified region of a particular bioagent is constrained in some fashion by its biological purpose (*i.e.*, RNA structure, protein coding, etc.). For example, protein coding regions are constrained by amino acid coding considerations, whereas a ribosome is mostly constrained by base pairing in stems and sequence constraints in unpaired loop regions. Moreover, different regions of the ribosome might have significant preferences that differ from each other.

[0117] One embodiment of application of the cloud algorithm is described in Example 1. By collecting all likely species amplicons from a primer set and enlarging the set to include all biologically likely variant amplicons using the cloud algorithm, a suitable cluster region of base count space is defined for a particular species of bioagent. The regions of base count space in which groups of related species are clustered are referred to as “bioclusters.”

[0118] When a biocluster is constructed, every base count in the biocluster region is assigned a percentage probability that a species variant will occur at that base count. To form a probability density distribution of the species over the biocluster region, the entire biocluster probability values are normalized to one. Thus, if a particular species is present in a sample, the probability of the species biocluster integrated over all of base count space is equal to one.

[0119] At this point in the ranking procedure, proposed target species to be detected are taken into account. These generally are the bioagents that are of primary importance in a particular detection scenario. For example, if *Yersinia pestis* (the causative agent of bubonic and pneumonic plague) were the target, the *Yersinia pestis* species biocluster identified as described above, would be the “target biocluster.” To complete the example, assume that all other database species serve as the scenario background. The discrimination metric in this case is defined as the sum total of all the biocluster overlap from other species into the *Yersinia pestis* biocluster.

[0120] In this example, the *Yersinia pestis* biocluster overlap is calculated as follows. A probability of detection of 99% ( $P_D=0.99$ ) is defined, although this value can be altered as needed. The “detection range” is defined as the set of biocluster base counts, of minimal number, that encloses 99% of the entire target biocluster. For each additional bacterial species in the database, the amount of biocluster probability density that resides in the base counts in the defined detection range is calculated and is the effective biocluster overlap between that background species and the target species. The sum of the biocluster overlap over all background species serves as the metric for measuring the discrimination ability of a defined target by a proposed primer set. Mathematically, because the most discriminating primer sets will have minimal biocluster overlap, an inverse figure of merit  $\phi$  is defined,  $1/\phi = \sum_{i=1}^N \theta_i$  where the sum is taken over the individual biocluster overlap values  $\theta_i$  from all  $N$  background species bioclusters ( $i=1, \dots, N$ ). For example, Figure 2 shows the inverse figure of merit  $\phi$  plotted for a master list of 16 primer sets using *Yersinia pestis* as the target biocluster. Using the inverse figure of merit minimization criteria defined above, the result is that primer set number 4 provides the best discrimination of any of the individual primer sets in the master list.

[0121] This set of discrimination criteria also can be applied to combinations of primer sets. The respective four-dimensional base count spaces from each primer set can be dimensionally concatenated to form a  $(4 \times N)$ -dimensional base count space for  $N$  primer sets. Nowhere in the biocluster definition is it necessary that the biocluster reside in a four-dimensional space, thus the biocluster analysis seamlessly adapts to any arbitrary dimensionality. As a result, a master list of primer sets can be searched and ranked according to the discrimination of any combination of primer sets with any arbitrary number of primer sets making up the combination.

[0122] Using again the example of *Yersinia pestis* as the target, improved discrimination is achieved through use of an increasing number of primers. For each number of primers value on

the x-axis, the plotted inverse figure of merit value is that obtained from the most discriminating group (that group with the minimum figure of merit for that number of primer sets simultaneously used for discrimination). The result is that after the best groups of 3 and 4 primer sets are found, the inverse figure of merit approaches one and goes no further. That means that there is the equivalent of one background species biocluster overlapping into the target biocluster. In this example it is the *Yersinia pseudotuberculosis* species biocluster, which cannot be discriminated from *Yersinia pestis* by any combination of the 16 primer sets in the example. Thus, using the “best” 3 or 4 primer sets in the master list, *Yersinia pestis* is essentially discriminated from all other species bioclusters.

[0123] Thus, on the one hand, probability clouds can be used to detect variants of known bioagents. On the other hand, this method of the present invention can be used to unambiguously determine that an unknown bioagent is not a likely variant of a known bioagent and at the same time, classify the bioagent in terms of similarity to the known bioagents in the database.

[0124] RNA viruses depend on an error-prone polymerase for replication and therefore their nucleotide sequences (and the resultant base compositions) drift over time within the functional constraints allowed by selection pressure. Base composition probability distribution of a viral species or group represents a probabilistic distribution of the above variations in the {A, G, C, and T} base composition space and can be derived by analyzing base compositions of all known isolates of that particular species.

[0125] In one embodiment of the invention, a model organism, such as the positive strand RNA virus, hepatitis C virus (HCV), can be used to model these sequence variations. Mutation probabilities can be derived from the observed variations among, *e.g.*, a number of HCV sequences. Table 1 below, lists mutation probabilities that were derived from the observed variations among 50 HCV-I b sequences. Six different regions within the genome of 120 nucleotide (nt) average length, were picked based on priming considerations and a maximum amplicon length criterion of -150 nt. Base composition probability distributions for a species were determined in two steps. In the first step, mutation probabilities, *i.e.*, the probabilities of occurrence of each type of substitution, insertion, or deletion, were derived by pairwise comparisons of all known HCV isolates in each target region, and an estimate of the maximum number of mutations that a sequence may undergo were calculated. In the second step, the

mutation probabilities and maxima derived from the model organism were used to estimate variations in base compositions for each test species and to calculate mutation probability distances ( $\Delta_m$ ) between the species in base composition space, which is calculated as the negative base 10 logarithm ( $-\log_{10} P$ ) of the cumulative probabilities of all possible mutations of the A, G C, and T base counts of one species that would lead to the other.

[0126] There are several approaches to classifying an unknown organism based on the base composition of certain amplicons. To illustrate these approaches, the classification technique for exemplary primer pairs is shown. The method can be applied to other primer pairs.

**Table 1. Position Independent, Nucleotide Mutation Probabilities Over 6 Training Sequences For HCV-1b**

| Mutation | Seq. 1  | Seq. 2 | Seq. 3 | Seq. 4 | Seq. 5 | Seq. 6 | All Seq. |
|----------|---------|--------|--------|--------|--------|--------|----------|
| A -> A   | 91.82%  | 88.42% | 91.98% | 92.51% | 91.08% | 89.89% | 93.30%   |
| A -> C   | 1.54%   | 1.22%  | 0.56%  | 2.25%  | 0.14%  | 0.61%  | 0.80%    |
| A -> G   | 6.28%   | 9.57%  | 7.16%  | 5.08%  | 8.52%  | 8.61%  | 5.59%    |
| A -> T/U | 0.36%   | 0.79%  | 0.30%  | 0.15%  | 0.26%  | 0.90%  | 0.30%    |
| A ->     | 0.00%   | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |
| C -> A   | 1.00%   | 0.64%  | 0.40%  | 1.29%  | 0.10%  | 0.22%  | 0.46%    |
| C -> C   | 89.91%  | 93.27% | 89.89% | 93.87% | 93.84% | 93.87% | 94.68%   |
| C -> G   | 1.26%   | 0.61%  | 0.76%  | 0.13%  | 0.00%  | 0.71%  | 0.37%    |
| C -> T/U | 7.83%   | 5.48%  | 8.95%  | 4.71%  | 6.06%  | 5.20%  | 4.49%    |
| C ->     | 0.00%   | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |
| G -> A   | 3.97%   | 6.93%  | 3.96%  | 4.29%  | 7.10%  | 2.52%  | 3.47%    |
| G -> C   | 1.22%   | 0.85%  | 0.60%  | 0.19%  | 0.00%  | 0.57%  | 0.41%    |
| G -> G   | 94.41%  | 91.93% | 95.29% | 94.96% | 92.72% | 96.77% | 95.93%   |
| G -> T/U | 0.41%   | 0.29%  | 0.15%  | 0.56%  | 0.18%  | 0.13%  | 0.19%    |
| G ->     | 0.00%   | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |
| T -> A   | 0.49%   | 0.77%  | 0.22%  | 0.21%  | 0.22%  | 0.58%  | 0.29%    |
| T -> C   | 16.21%  | 10.23% | 9.61%  | 11.40% | 7.68%  | 9.17%  | 7.67%    |
| T -> G   | 0.88%   | 0.39%  | 0.20%  | 0.93%  | 0.18%  | 0.30%  | 0.30%    |
| T -> T/U | 82.42%  | 88.61% | 89.96% | 87.46% | 91.92% | 89.95% | 91.75%   |
| T ->     | 0.00%   | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |
| -> A     | 0.00%   | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |
| -> C     | 0.00%   | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |
| -> G     | 0.00%   | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |
| -> T/U   | 0.00%   | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |
| Total -> | 100.00% | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%    |

[0127] There are several approaches to classifying an unknown organism based on the base composition of certain amplicons. To illustrate these approaches, the classification technique for exemplary primer pairs is shown. The method can be applied to other primer pairs.

**[0128]** To develop a pattern classifier, the known base composition counts of amplicons of known organisms are used to construct the pattern classifier as a training set. In one embodiment of the pattern classifier, for each pattern class a base organism serves as a central point. For that pattern class, a distance is calculated from each organism in the training set to the base organism. The maximum distance found in this manner defines the class within the pattern classifier; all organisms less than the maximum distance to the base organism fall within the class.

**[0129]** Once the pattern classifier has been trained the unknown organism can be classified by determining the distance between the unknown organism and the base organism for each pattern. If the unknown organism falls within the maximum distance determined in the training process, the organism is classified as belonging to the same pattern class as the base organism. If the unknown organism falls outside the maximum distance, a probability that the organism belongs to the class can be derived as a function of the distance from the unknown organism to the base organism.

**[0130]** In an alternate embodiment of the pattern classifier, rather than identifying a base organism, a pattern is defined by selecting a centroid, which may not correspond to an actual organism, but serves as a center for the pattern class. During the training process, the centroid and the maximum distance is determined. Once trained, the classification of an unknown organism follows much the same as described above.

**[0131]** Several criteria for measuring the distance between organisms can be employed. For a particular primer-pair, the distance between the base compositions can be used. That is, if the base counts are treated as a mathematical vector, the distance between the vectors is the measure of distance.

**[0132]** As an example, the 229 E Human Coronavirus has a base count in the RdRp target region of A25,G24,C11,T28 and the SARS Coronavirus has a base count of A27,G19,C14,T28. Using the first example of distance (a Euclidean distance), the distance between them is 6.164.

**[0133]** An alternative measure of distance is to use the probability of mutation to derive distance. There are a number of mutation pathways between two polynucleotide sequences, which comprises a series of one or more mutation events. Based on empirical finding, the probability of individual mutations is known. Table 1 shows a list of typical individual

mutations with their associated probabilities. The probability of a specific mutation pathway is the product of the probabilities of the individual mutations. One method of defining distance is to take the sum of all probabilities of all mutations pathways,  $P$ . The mutational distance between the two polynucleotide sequences can be defined as  $-\log_{10} P$ . In the above example, the distance between the 229E Human Coronavirus and the SARS coronavirus is 8.8. It should be noted that since longer mutation pathways are less likely, only certain mutations are needed to get from 229E to SARS, and thus the longer pathways can be discarded.

[0134] Figure 3 is a graph showing the base compositions of the 229E Human Coronavirus, OC43 Human Coronavirus and the SARS Coronavirus. In this graph, the A, G, and C base counts are plotted on the axes and the T base count is represented by using rotation.

[0135] Figure 4 shows a number of animal coronavirus species. The branches on the tree represent the phylogenetic relationship between the various taxons. For each taxonomic grouping, an oval represents the maximal distance between any two members of the group represented by  $\Delta_m$  next to the oval. For example, the bovine isolates (BCoV-Quebec and BCov-Lun) are clustered together ( $\Delta_m < 2.0$ ), and are closer to each other than to their nearest neighbor on the phylogenetic tree, HCoV-OC43. The bovine and the OC43 species form a closely related cluster with a relatively high probability of misclassification ( $(\Delta_m < 4.5)$ ). Similarly, the murine and rat coronavirus isolates are closely related species that can not be distinguished from each other using just two target regions ( $\Delta_m < .9$ ), yet the rodent viruses are easily distinguished from the bovine/OC43 group ( $\Delta_m < 6.8$ ). Similarly, many of the group 1 animal coronaviruses (CCoV, FCoV, TGEV) clustered together and were very close to each other in mutation and base composition distance. These, therefore, could potentially be misclassified at the species level ( $\Delta_m < 4.7$ ). This is consistent with previous reports that suggest that CCoV are serologically and genetically related to other group 1 animal coronaviruses. However, this group was clearly resolved from other members of group 1 coronaviruses such as 229E and PEDV ( $\Delta_m < 11.6$ ). In contrast to the group 1 and group 2 species clusters, the two target regions chosen here did not cluster the group 3 species together. The three known isolates of avian coronaviruses were as far away from each other as they were from members of group 1 coronaviruses. Overall, the mutation-distance analysis suggests that the previously known members of group 2 coronaviruses represent a clearly delineated group, well resolved from groups 1 and 3. In contrast, no clear delineation between groups 1 and 3 was observed.

[0136] Further refinement to the classification can be made by assigning a match probability of an unknown for each pattern class by calculating the distance to each pattern class. By applying additional pattern classifiers based on other primer pairs, the ability to resolve unknowns is enhanced. In the example described above, it would be difficult to distinguish an unknown in group 1 from group 3 for the given primer pair. Applying the pattern classifier with other primer pairs may yield a greater distance between group 1 and group 3 coronaviruses. This triangulation approach is described further below.

[0137] In alternate embodiments of the mutational probability model, a centroid is not chosen and restrictions among strains were compared to one another. Using best estimates of the phylogenetic tree, only descendants were compared to their direct forebears, for a direct estimate of a mutational probability. This comparison had the effect of reducing the magnitude of the mutation probabilities.

[0138] Because it is known that DNA triplets code for a single amino acid, in some embodiments, for primer regions that are in a protein-coding region of the sequence, the mutational probabilities are determined in a position-dependent way, so that the 20 types of mutations (12 substitutions, 4 deletions, and 4 insertions) are now expanded to a set of 60 (20 types x 3 positions). It is well known that the first position of a triplet is highly conserved, while the third position is the least conserved (and it is referred to as a wobble position because of this) and this is reflected in the different mutation probabilities per position.

[0139] In other embodiments, the mutational probability model incorporates both the restrictions among strains and position dependence of a given nucleobase within a triplet. In one embodiment of the invention, a polytope pattern classifier is used to classify test or unknown organism according to its amplicon base composition. The polytope pattern classifier of the present invention defines the bounds of a pattern class by a convex polytope. The polytope pattern classifier is trained by defining a minimal polytope which contains all the samples in the training set.

[0140] Generally, a polytope can be expressed by a system of linear inequalities. Data supplied to the pattern classifier are typically expressed as an  $n$ -dimensional vector. Accordingly, an  $n$ -dimensional polytope can be expressed as a system of inequalities of the form:

$$a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n \leq C$$

and of the form,

$$D \leq b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n.$$

According to one embodiment of the present invention, the components of the data vectors are integers. Thus, the polytopes can be reduced to a system of linear inequalities of the following form,

$$D \leq a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n \leq C, \text{ where each } a_i \text{ is either 0 or 1.}$$

To define a minimal polytope, all inequalities of the form equation shown above can be used for all combinations of  $a_i$ . During the training process the constants  $C$  and  $D$  are determined for each inequality.

[0141] In certain aspects of the invention, a density is defined for each polytope by taking the total number of samples in the training set residing in the polytope and dividing by the total volume of the polytope. Once the polytopes are calculated for each pattern class identified in the training set, the polytope pattern classifier is trained and can be applied to test or unknown data. In classifying an unknown represented by a data vector, the distance to each pattern class is calculated. A point density of the data vector to a polytope is defined to be the density of the polytope multiplied by a decay factor which is a function of distance of the data vector to the polytope. A match probability to each of the classes is calculated based on the point density. In one embodiment of the invention, for example, the match probability can be the normalized average of all point densities for that particular data sample.

[0142] It should be noted that the measure of volume and distance described in the density and point density calculations need not be standard Euclidean-based measures of distance and volume. For example, if the data vectors have integer components, the volume of a polytope can be defined as a lattice volume that is the number of integer lattice points within a given polytope. Similarly, the distance from a point to a polytope can be defined as a lattice distance that is the minimum number of lattice points traversed between a point and any point within the polytope.

[0143] Figure 5A is a flow chart illustrating a method of training an embodiment of a polytope pattern classifier. At step 1202, a training sample is received from a training set. Associated with each training sample is the pattern class it is a member of. At step 1204, the pattern class is determined. At step 1206, if necessary that pattern class' polytope is modified so as to incorporate the training sample. If the training sample lies within the current version of the pattern class' polytope, no modification is required. This modification typically takes the form of

comparing the training sample to the existing inequalities that defined the polytope. If the training sample falls outside an inequality, the inequality is modified to incorporate the training sample. In the modification process, the inequality is modified to expand the polytope as little as possible. At step **1208**, the process iterates to the next training sample, if any remain. Otherwise, the training is complete.

**[0144]** One should note that though the flowchart describes an iteration through the training samples and in polytope modification, an iteration through the inequalities which defined the polytope, the order of iteration could be equivalently transposed. That is, rather than considering each training sample first, each inequality is considered. For each inequality, the training sample is compared against the inequality and the inequality is modified to accommodate the training sample if necessary. Then the iteration can continue to the next inequality.

**[0145]** Figure 5B is a flow chart illustrating the method of identifying an unknown sample using an embodiment of a trained polytope pattern classifier. At step **1222**, an unknown sample is received by the polytope pattern classifier. At step **1224**, a pattern class is selected. At step **1226**, the distance between the pattern class' polytope and the unknown sample is calculated. Based on the distance, at step **1228**, the point density of the unknown sample with respect to the pattern class is calculated. At step **1230**, the process repeats for the next pattern class. When all point densities with respect to all the pattern classes are calculated, a match probability is generated by normalizing the point densities at step **1232**.

**[0146]** To simplify the complexity of higher dimensional polytope pattern classifiers, a plurality of lower dimensional polytope pattern classifiers can be used. According to this embodiment of the invention, all data including unknowns and the data in the training set, are divided into a plurality of subspaces having the lower dimension. A polytope pattern classifier is associated with each subspace. Each polytope pattern classifier is trained on the subset of the training set that resides within the associated subspace. Once trained, the one of the plurality of subspaces to which an unknown belongs is first applied, then the polytope pattern classifier associated with that subspace is applied to the data.

**[0147]** In certain aspects of the invention, subspaces are defined by the length of the data, *e.g.* the amplicon length. When the components of the data vectors are integers, the subspaces determined in this manner can yield a finite if not small number of subspaces.

[0148] In an alternative embodiment of the polytope pattern classifier, contributions from all polytopes are considered, regardless of which subspace the unknown data belongs to. For example, the point density of an unknown to a given pattern class can be a function of the distance of the unknown data vector to every polytope associated with a given pattern class. In order to simplify this calculation, the distance can be broken into two components, the distance between the unknown data vector to the subspace containing the polytope, and the distance between a projected data vector, *i.e.*, the data vector when projected onto the subspace containing the polytope, and the polytope. These two components of the data vector can be into different decay factors.

[0149] Figure 6A is a flow chart illustrating the method of training an embodiment of a polytope pattern classifier of a lower dimension when the sample space is reduced in dimension by imposing a constraint. At step 1302, a training sample is received from the training set. The constraint is applied to determine which subspace the training sample belongs to at step 1304. The training sample is placed into a training subset corresponding to that subspace, at step 1306. At step 1308, the process is made to repeat, until all training samples have been grouped into corresponding subspaces. Then at step 1310, a subspace is selected along with the corresponding subset of the training samples. At step 1312 the pattern classifier corresponding to that subspace is trained. It can be trained using a method like that described in Figure 6A. At step 1314, the process is made to repeat, until all subspaces derived from the constraint have fully trained pattern classifiers. It should be noted that in another method of training the order can be changed. For example, after the subspace of a training sample is identified, it can be used to train the corresponding pattern classifier immediately rather than waiting until all training samples are sorted. The flow chart is intended to clearly describe an example of a training method.

[0150] Figure 6B is a flow chart illustrating a method of identifying a unknown sample in a manner similar to that of Figure 5B. At step 1332, an unknown sample is received by the pattern classification system. At step 1334, the constraint is applied and the subspace to which the sample belongs is determined. Steps 1336, 1338, 1340, 1342, and 1344 apply a similar same pattern identification algorithm to that described in steps 1224, 1226, 1228, 1230, and 1232 respectively, where the polytope associated with each pattern class used is the polytope contained in the subspace to which the sample belongs. It should be noted that depending on the members of the various pattern classes, a pattern class can have more than one polytope, but in different subspaces.

[0151] The method described in Figure 6B does not account for the polytopes for a given class in subspaces other than that to which the sample belongs. Figure 6B is a flow chart illustrating an alternative method of identifying an unknown sample using polytope classifiers trained by a process such as that described in Figure 6A. At step 1352, an unknown sample is received by the pattern classification system. At step 1354, a pattern class is selected. At step 1356, a subspace is selected which contains one of the pattern class' polytopes. If no polytope for that pattern class exists in that subspace, another subspace can be selected. At step 1358, a gap distance is calculated, *i.e.* the distance between the unknown sample and the selected subspace. At step 1360, the mutation distance is calculated, *i.e.* the distance between a "projection" of the unknown sample and the pattern class' polytopes. In practice, the distance is actually the minimum distance between all possible minimal insertions (or deletions) sufficient to mutate the sample to the given subspace. At step 1362, the point density of the unknown sample with respect to the pattern class' polytope is calculated as a function of either the gap distance, the mutation distance or both. At step 1364, the process is made to repeat until all subspaces with the specific pattern class' polytopes have been selected. Once all the point densities have been calculated, at step 1366, the point probabilities are all combined to produce a composite point probability for the unknown sample with respect to the entire pattern class. At step 1368, the process is made to repeat until all pattern classes have been selected. When all point densities with respect to all the pattern classes are calculated, a match probability is generated by normalizing the point densities at step 1370.

[0152] Specifically, as applied to the classification of an unknown organism, the polytope pattern classifier is applied to data vectors representing the amplicon base composition of organisms. The polytope pattern classifiers are trained on the amplicon base compositions of known organisms using a database of known organism amplicon mass spectra that has been indexed for key parameters of amplicon DNA sequence, including amplicon length, base composition and ratios of key nucleotides (*e.g.*, C + T, G +T, G+C). In one aspect of the invention, the amplicon database is organized according to taxonomic identification of the known organisms. In certain aspects of the invention, the database includes amplicon data for all known organisms in a given genus, order, class, phyla, or kingdom.

[0153] In one embodiment of the present invention, each amplicon is analyzed separately. For each amplicon, a taxon is associated with at least one pattern class. When considering a given amplicon, the data used in classification lies within the theoretical maximum base composition

space defined by the content of A, G, C and T bases. Thus, the data used in classification can be represented by a four dimensional vector. Furthermore, these base counts result in integer values.

[0154] To further simplify the classifier models, the data are subdivided into potential pattern subclasses based on amplicon length. By applying a constraint to the length of the data vectors, three dimensional pattern classifiers can be employed.

[0155] For example, Table 2 (below) represents a set of known organisms belonging to the *Neisseriales* taxon. The base compositions for bioagent identifying amplicons obtained with a broad range bacterial primer pair are shown. Within the known taxons of *Neisseriales*, for example, the amplicons are either 55 or 56 nucleotides in length. In accordance with the use of three dimensional polytope classification, the data are broken into two groups where each member has the same amplicon length. For illustrative purposes, the training of a three-dimensional classifier on a training set comprising data of amplicon length 56 is considered. In the figures, the polyhedra (3-dimensional polytopes) are shown in the G, C, and T axis. First unary inequalities are applied to first define the polyhedron, these inequalities are derived selecting a smallest unary inequality ranges for which the data in the training sets still reside within the polyhedron. For the given example, these inequalities are  $16 \leq G \leq 18$ ,  $13 \leq C \leq 16$ , and  $7 \leq T \leq 11$ . As illustrated in Figure 7A, these inequalities define a polyhedron of volume 60. It should be noted that the A composition value was not used since the value of A is governed by the amplicon length. However, it should be noted that from the training set, a minimal unary inequality of  $15 \leq A \leq 17$  can be derived. Because of the constraint on amplicon length, this is equivalent to the trinary inequality of  $39 \leq G+C+T \leq 41$ . Figure 7B shows the result of boundaries of this inequality and Figure 7C shows the resultant polyhedron when the inequality is applied, resulting in a polyhedron with the volume of 31.

**Table 2. *Neisseriales* Base Compositions for a Representative  
Broad Range Bacterial Survey Primer Pair**

| Bioagent                            | Base Composition |    |    |    |         |
|-------------------------------------|------------------|----|----|----|---------|
|                                     | A                | G  | C  | T  | A+G+C+T |
| <i>Neisseria gonorrhoeae</i> FA1090 | 16               | 16 | 13 | 10 | 55      |
| <i>Neisseria meningitidis</i> A     | 16               | 16 | 15 | 8  | 55      |
| <i>Neisseria meningitidis</i> B     | 16               | 16 | 15 | 8  | 55      |
| <i>Neisseria meningitidis</i> C     | 16               | 16 | 15 | 8  | 55      |
| <i>Chromobacterium violaceum</i>    | 16               | 18 | 15 | 6  | 55      |
| <i>Neisseria gonorrhoeae</i> B 5025 | 16               | 16 | 13 | 11 | 56      |
| <i>Neisseria weaveri</i>            | 16               | 16 | 13 | 11 | 56      |
| <i>Formivibrio citricus</i>         | 17               | 16 | 16 | 7  | 56      |

|                                      |    |    |    |   |    |
|--------------------------------------|----|----|----|---|----|
| <i>Aquaspirillum delicatum</i>       | 15 | 17 | 15 | 9 | 56 |
| <i>Aquaspirillum sinuosum</i>        | 15 | 17 | 15 | 9 | 56 |
| <i>Aquaspirillum gracile</i>         | 15 | 17 | 16 | 8 | 56 |
| <i>Microvigula aerodenitrificans</i> | 16 | 18 | 14 | 8 | 56 |

[0156] In addition, individual binary inequalities can be applied. While within the A, G, C, T space, there are six possible binary inequalities, there are only three in the G, C, T space as the binary inequalities involving A are accounted for because of the constraint on amplicon length.

[0157] Figure 8A illustrates the application of the  $22 \leq C+T \leq 24$  binary inequality and shows the boundaries imposed by the inequality to the existing polyhedron. Figure 8B shows the resultant polyhedron, which has a volume of 26. This inequality is a constraint on the composition of purines (C+T) in the amplicons determined. As will be apparent to the skilled artisan, constraining the polyhedron according to pyrimidine composition can be considered complementary to the purine constraint, because of the constraints on amplicon length. Figures 9A and 9B show the result of applying the keto/amino preference (G+T binary inequality). Figure 10 shows the result of applying the strong/weak base pairing constraints (G+C binary inequality). In this example, the resulting polyhedral pattern class is reduced to a minimum volume of 23.

[0158] A density calculation can also be performed based on the number of amplicons that occupy the taxon. For this example, the 7 amplicons occupy a volume of 23 in base compositional space giving a density of 0.304.

[0159] Though not shown, similar classification training results a pattern classifier where the amplicons of length 55 generate a polyhedron of volume 9. With 5 exemplars in the training set, a density of 0.556 can be calculated.

[0160] The skilled artisan will recognize that the polytopes thus generated can be generated or represented in various forms, including but not limited to, 4 dimensions rather than 3, and the minimum volume of base compositions space may be observed by varying the parameters used to constrain the polyhedrons.

[0161] For a given amplicon length, in one embodiment of the invention, the multidimensional polyhedron space and the density thereof can be determined for all taxonomic groups. As shown

in Figures 11A-E, the polyhedrons for each individual taxon can be superimposed, while the constraints imposed by the sum of all the taxons in, for example, a given class can be independently applied to define the overall base compositional space occupied. It will be apparent to the skilled artisan that the polyhedrons for each taxon may overlap, while the overall base compositional space of the larger class taxon may occupy space for which no model organism has been observed (Figure 11F).

[0162] Shown in Figure 11F, an unknown bioagent is determined to have a 346 base composition of A=15, G=18, C=16, T=7, which has a total length of 56. Accordingly, the polytope pattern classifier trained on amplicons of length 56 is used. As shown in Figure 11F, the base composition resides in the polytope for the *Birkholderiales* Taxon and *Hydrogenophilales* Taxon and has a distance of 1 (determined by lattice hops) to the remain taxons. The point densities for each taxon are determined by applying a decay factor of  $1/256$  raised to the power of the distance. The resultant match probabilities are then calculated by normalizing the point densities. In the example only 5 bacteriological orders are shown, but the results are normalized to all 71 bacteriological orders, but most are not shown for clarity.

[0163] In an alternate embodiment of the pattern classifier, the point densities can be calculated by combining the density values derived from polytopes all representing a specific taxon. In the example shown above, the *Neisseriales* pattern class comprises amplicons of both length 55 and length 56, as a result in the training of the pattern classifier there is a polytope in the “55 length subspace” associated with the *Neisseriales* pattern class (henceforth the *Niesserales-55* polytope) and a polytope in the “56 length subspace” also associated with the *Neisseriales* pattern class (henceforth the *Niesserales-56* polytope). The alternate pattern classifier uses both polytopes for identification of the unknown sample. In the preceding example, there is a distance of 1 between the unknown sample and the *Neisseriales-56*. In deriving the distance between the unknown sample and the *Niesserales-55* polytope, the distance measure can be broken into two distance components, the distance between the sample and the “55 length subspace” which is 1 and the distance between the sample projected onto the 55 length subspace to the *Neisseriales-55* polytope is 1. The first component of distance is referred to as the “gap distance” and the second component of the distance is referred to as the “mutation distance.” In this case, the projection is the point in the 55 length subspace which lies closest to the *Neisseriales-55* polytope with only one change in A, G, C, or T. If the gap distance were 2, the projection would be the point in the subspace which lies closest to the polytope have

- 44 -

at most two changes in A, G, C, or T. It should be noted that since the unknown sample resides in the 56 length subspace, the gap distance between the unknown sample and the *Neisseriales*-56 polytope is 0.

[0164] However, the match probability based on a single primer pair may not provide accurate results. According to the present invention, the assignment of an unknown bioagent to a taxon can be further refined by comparing the base compositional space occupied by additional amplicons (Figure 12). Using this “triangulation” approach, the normalized product of the individual primer pair probabilities yields a global assignment probability for each taxon. Thus, in certain embodiments of the invention, an unknown bioagent is matched in base compositional space to the 1, 2, 3, 4 or more polyhedrons representing the base compositional space of different amplicons from known bioagents (the “training set”).

[0165] Probability calculations can be applied to determine reliability of the method, as summarized in Table 3 below, wherein the primer pair numbers refer to primer pairs disclosed in commonly owned U.S. application Serial No. : 11/060,135.

**Table 3. Reliability of Taxonomic Assignment of Bacteria using the Polytope Pattern Model.**

| Assign-<br>ment<br>Threshold | Primer<br>Pair<br>Comb. | % of assignment above threshold |       |       |        |       | % of correct assignment |       |       |        |       |
|------------------------------|-------------------------|---------------------------------|-------|-------|--------|-------|-------------------------|-------|-------|--------|-------|
|                              |                         | Phylum                          | Class | Order | Family | Genus | Phylum                  | Class | Order | Family | Genus |
| 50%                          | 346                     | 48.6%                           | 32.8% | 32.4% | 33.1%  | 31.7% | 70.6%                   | 70.0% | 67.6% | 60.4%  | 57.1% |
|                              | 347                     | 86.2%                           | 79.8% | 65.2% | 61.7%  | 56.6% | 84.8%                   | 73.0% | 74.3% | 70.7%  | 71.3% |
|                              | 348                     | 92.4%                           | 71.6% | 66.4% | 62.4%  | 65.3% | 79.9%                   | 82.4% | 78.2% | 73.8%  | 76.0% |
|                              | 361                     | 97.1%                           | 97.4% | 97.4% | 97.9%  | 95.9% | 87.7%                   | 94.7% | 87.3% | 83.6%  | 75.2% |
|                              | 346 +<br>347            | 85.7 %                          | 77.4% | 79.3% | 80.9%  | 80.3% | 87.1%                   | 91.1% | 83.9% | 88.3%  | 85.2% |
|                              | 346 +<br>348            | 96.4%                           | 82.8% | 86.4% | 88.1%  | 85.3% | 83.5%                   | 91.0% | 82.8% | 83.0%  | 83.8% |
|                              | 346 +<br>361            | 87.6%                           | 64.5% | 71.4% | 73.3%  | 75.5% | 81.1%                   | 87.4% | 85.5% | 80.9%  | 84.0% |

|                                |       |       |       |       |       |       |       |       |       |       |
|--------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 347 +<br>348                   | 97.2% | 94.7% | 93.6% | 91.7% | 91.0% | 90.4% | 92.2% | 89.7% | 89.1% | 86.9% |
| 347 +<br>361                   | 92.8% | 89.3% | 90.7% | 84.7% | 86.0% | 91.1% | 91.9% | 87.1% | 87.8% | 83.0% |
| 348 +<br>361                   | 96.9% | 86.7% | 84.5% | 82.9% | 87.9% | 85.1% | 94.6% | 87.8% | 85.4% | 85.7% |
| 346 +<br>347 +<br>348          | 94.1% | 92.9% | 92.9% | 95.0% | 92.9% | 89.6% | 95.2% | 91.3% | 90.9% | 86.6% |
| 346 +<br>347 +<br>361          | 90.5% | 87.9% | 89.0% | 90.5% | 89.3% | 90.9% | 94.5% | 90.1% | 92.8% | 89.6% |
| 346 +<br>348 +<br>361          | 95.7% | 87.4% | 87.4% | 91.9% | 89.7% | 87.0% | 95.7% | 91.9% | 88.9% | 89.2% |
| 347 +<br>348 +<br>361          | 97.8% | 94.7% | 92.8% | 95.9% | 94.0% | 93.5% | 96.5% | 92.8% | 91.7% | 90.8% |
| 346 +<br>347 +<br>348 +<br>361 | 95.9% | 95.5% | 93.3% | 96.0% | 92.8% | 89.4% | 96.6% | 93.2% | 94.3% | 91.4% |
| 60%                            | 88.4% | 88.8% | 88.1% | 91.6% | 88.6% | 94.3% | 97.5% | 96.3% | 95.5% | 93.2% |
| 70%                            | 81.7% | 81.9% | 82.1% | 86.2% | 84.5% | 96.8% | 97.9% | 96.8% | 95.8% | 94.1% |
| 80% 346 +<br>347 +             | 66.9% | 72.2% | 76.0% | 81.6% | 77.4% | 97.9% | 98.6% | 98.0% | 96.8% | 96.0% |
| 90% 348 +<br>361               | 55.3% | 61.2% | 66.6% | 69.1% | 70.7% | 99.1% | 99.2% | 98.7% | 98.0% | 96.1% |

[0166] Table 3 provides a summary of the polytope analysis of 580 test bioagents (sample set) compared to 3413 individual known species in the training set. To date, 14/19 Phyla, 22/28 Classes, 56/71 Orders, 119/170 Families, 229/466 Genera have been analyzed. Figure 13 illustrates that reliable phylogenetic assignment can be made using the polytope pattern model. In certain embodiments of the invention alternate compatible assignments may be suggested.

The present invention contemplates that in some circumstances the present invention will generate multiple possible phylogenetic assignments in parallel at different levels, allowing at least a partial assignment of unknown bioagents.

[0167] In some embodiments, the identity and quantity of an unknown bioagent can be determined using the process illustrated in Figure 14. Primers (500) and a known quantity of a calibration polynucleotide (505) are added to a sample containing nucleic acid of an unknown bioagent. The total nucleic acid in the sample is then subjected to an amplification reaction (510) to obtain amplification products. The molecular masses of amplification products are determined (515) from which are obtained molecular mass and abundance data. The molecular mass of the bioagent identifying amplicon (520) provides the means for its identification (525) and the molecular mass of the calibration amplicon obtained from the calibration polynucleotide (530) provides the means for its identification (535). The abundance data of the bioagent identifying amplicon is recorded (540) and the abundance data for the calibration data is recorded (545), both of which are used in a calculation (550) which determines the quantity of unknown bioagent in the sample.

[0168] A sample comprising an unknown bioagent is contacted with a pair of primers which provide the means for amplification of nucleic acid from the bioagent, and a known quantity of a polynucleotide that comprises a calibration sequence. The nucleic acids of the bioagent and of the calibration sequence are amplified and the rate of amplification is reasonably assumed to be similar for the nucleic acid of the bioagent and of the calibration sequence. The amplification reaction then produces two amplification products: a bioagent identifying amplicon and a calibration amplicon. The bioagent identifying amplicon and the calibration amplicon should be distinguishable by molecular mass while being amplified at essentially the same rate. Effecting differential molecular masses can be accomplished by choosing as a calibration sequence, a representative bioagent identifying amplicon (from a specific species of bioagent) and performing, for example, a 2-8 nucleobase deletion or insertion within the variable region between the two priming sites. The amplified sample containing the bioagent identifying amplicon and the calibration amplicon is then subjected to molecular mass analysis by mass spectrometry, for example. The resulting molecular mass analysis of the nucleic acid of the bioagent and of the calibration sequence provides molecular mass data and abundance data for the nucleic acid of the bioagent and of the calibration sequence. The molecular mass data obtained for the nucleic acid of the bioagent enables identification of the unknown bioagent and

the abundance data enables calculation of the quantity of the bioagent, based on the knowledge of the quantity of calibration polynucleotide contacted with the sample.

[0169] In some embodiments, construction of a standard curve where the amount of calibration polynucleotide spiked into the sample is varied, provides additional resolution and improved confidence for the determination of the quantity of bioagent in the sample. The use of standard curves for analytical determination of molecular quantities is well known to one with ordinary skill and can be performed without undue experimentation.

[0170] In some embodiments, multiplex amplification is performed where multiple bioagent identifying amplicons are amplified with multiple primer pairs which also amplify the corresponding standard calibration sequences. In this or other embodiments, the standard calibration sequences are optionally included within a single vector which functions as the calibration polynucleotide. Multiplex amplification methods are well known to those with ordinary skill and can be performed without undue experimentation.

[0171] In some embodiments, the calibrant polynucleotide is used as an internal positive control to confirm that amplification conditions and subsequent analysis steps are successful in producing a measurable amplicon. Even in the absence of copies of the genome of a bioagent, the calibration polynucleotide should give rise to a calibration amplicon. Failure to produce a measurable calibration amplicon indicates a failure of amplification or subsequent analysis step such as amplicon purification or molecular mass determination. Reaching a conclusion that such failures have occurred is in itself, a useful event.

[0172] In some embodiments, the calibration sequence is comprised of DNA. In some embodiments, the calibration sequence is comprised of RNA.

[0173] In some embodiments, the calibration sequence is inserted into a vector which then itself functions as the calibration polynucleotide. In some embodiments, more than one calibration sequence is inserted into the vector that functions as the calibration polynucleotide. Such a calibration polynucleotide is herein termed a "combination calibration polynucleotide." The process of inserting polynucleotides into vectors is routine to those skilled in the art and can be accomplished without undue experimentation. Thus, it should be recognized that the calibration method should not be limited to the embodiments described herein. The calibration

method can be applied for determination of the quantity of any bioagent identifying amplicon when an appropriate standard calibrant polynucleotide sequence is designed and used. The process of choosing an appropriate vector for insertion of a calibrant is also a routine operation that can be accomplished by one with ordinary skill without undue experimentation.

[0174] Bioagents that can be identified by the methods of the present invention include RNA viruses. The genomes of RNA viruses can be positive-sense single-stranded RNA, negative-sense single-stranded RNA or double-stranded RNA. Examples of RNA viruses with positive-sense single-stranded genomes include, but are not limited to members of the Caliciviridae, Picornaviridae, Flaviviridae, Togaviridae, Retroviridae and Coronaviridae families. Examples of RNA viruses with negative-sense single-stranded RNA genomes include, but are not limited to, members of the Filoviridae, Rhabdoviridae, Bunyaviridae, Orthomyxoviridae, Paramyxoviridae and Arenaviridae families. Examples of RNA viruses with double-stranded RNA genomes include, but are not limited to, members of the Reoviridae and Birnaviridae families.

[0175] In some embodiments of the present invention, RNA viruses are identified by first obtaining RNA from an RNA virus, or a sample containing or suspected of containing an RNA virus, obtaining corresponding DNA from the RNA by reverse transcription, amplifying the DNA to obtain one or more amplification products using one or more pairs of oligonucleotide primers that bind to conserved regions of the RNA viral genome, which flank a variable region of the genome, determining the molecular mass or base composition of the one or more amplification products and comparing the molecular masses or base compositions with calculated or experimentally determined molecular masses or base compositions of known RNA viruses, wherein at least one match identifies the RNA virus. Methods of isolating RNA from RNA viruses and/or samples containing RNA viruses, and reverse transcribing RNA to DNA are well known to those of skill in the art.

[0176] Members of the Filoviridae, Flaviviridae, Bunyaviridae and Arenaviridae families represent RNA virus examples of bioagents which can be identified by the methods of the present invention. Filoviruses, flaviviruses, arenaviruses and three genera of the Bunyaviridae family (hantavirus, phlebovirus and nairovirus) are known to cause to VHF.

[0177] In one embodiment of the present invention, the target gene is filovirus RNA-dependent RNA polymerase. In another embodiment, the target gene is filovirus nucleocapsid.

[0178] In one embodiment of the present invention, the target gene is flavivirus NS5, the viral RNA-dependent RNA polymerase. In another embodiment, the target gene is flavivirus NS3, the viral protease, helicase and NTPase.

[0179] In one embodiment of the present invention, the target gene is hantavirus RNA-dependent RNA polymerase. In another embodiment, the target gene is hantavirus nucleocapsid. In another embodiment, the target gene is phlebovirus RNA-dependent RNA polymerase. In another embodiment, the target gene is nairovirus nucleocapsid.

[0180] In one embodiment of the present invention, the target gene is the arenavirus gene L, which is the viral RNA-dependent RNA polymerase. In another embodiment, the target gene is arenavirus NP, the viral nucleocapsid.

[0181] In other embodiments of the present invention, the intelligent primers produce bioagent identifying amplicons within stable and highly conserved regions of hantaviral, phleboviral or nairoviral genomes. The advantage to characterization of an amplicon in a highly conserved region is that there is a low probability that the region will evolve past the point of primer recognition, in which case, the amplification step would fail. Such a primer set is thus useful as a broad range survey-type primer. In another embodiment of the present invention, the intelligent primers produce bioagent identifying amplicons in a region which evolves more quickly than the stable region described above. The advantage of characterization bioagent identifying amplicon corresponding to an evolving genomic region is that it is useful for distinguishing emerging strain variants.

[0182] The present invention also has significant advantages as a platform for identification of diseases caused by emerging viruses. The present invention eliminates the need for prior knowledge of bioagent sequence to generate hybridization probes. Thus, in another embodiment, the present invention provides a means of determining the etiology of a virus infection when the process of identification of viruses is carried out in a clinical setting and, even when the virus is a new species never observed before. This is possible because the methods are not confounded by naturally occurring evolutionary variations (a major concern for characterization of viruses which evolve rapidly) occurring in the sequence acting as the template for production of the bioagent identifying amplicon. Measurement of molecular mass and determination of base composition is accomplished in an unbiased manner without sequence prejudice.

- 50 -

[0183] Another embodiment of the present invention also provides a means of tracking the spread of any species or strain of virus when a plurality of samples obtained from different locations are analyzed by the methods described above in an epidemiological setting. In one embodiment, a plurality of samples from a plurality of different locations are analyzed with primers which produce bioagent identifying amplicons, a subset of which contain a specific virus. The corresponding locations of the members of the virus-containing subset indicate the spread of the specific virus to the corresponding locations.

[0184] The present invention also provides kits for carrying out the methods described herein. In some embodiments, the kit may comprise a sufficient quantity of one or more primer pairs to perform an amplification reaction on a target polynucleotide from a bioagent to form a bioagent identifying amplicon. In some embodiments, the kit may comprise from one to fifty primer pairs, from one to twenty primer pairs, from one to ten primer pairs, or from two to five primer pairs. In some embodiments, the kit may comprise one or more primer pairs recited in Tables 4-7.

[0185] In some embodiments, the kit may comprise one or more broad range survey primer(s), division wide primer(s), or drill-down primer(s), or any combination thereof. A kit may be designed so as to comprise particular primer pairs for identification of a particular bioagent. For example, a broad range survey primer kit may be used initially to identify an unknown bioagent as a member of the filovirus genus. Another example of a division-wide kit may be used to distinguish Zaire Ebola virus, Sudan Ebola virus and Marburg virus from each other. A drill-down kit may be used, for example, to distinguish different subtypes of Zaire Ebola virus, or to identify genetically engineered filoviruses. In some embodiments, any of these kits may be combined to comprise a combination of broad range survey primers and division-wide primers so as to be able to identify the species of an unknown bioagent.

[0186] In some embodiments, the kit may contain standardized calibration polynucleotides for use as internal amplification calibrants. Internal calibrants are described in commonly owned U.S. Patent Application Serial No: 60/545,425.

[0187] In some embodiments, the kit may also comprise a sufficient quantity of reverse transcriptase (if an RNA virus is to be identified for example), a DNA polymerase, suitable nucleoside triphosphates (including any of those described above), a DNA ligase, and/or reaction

- 51 -

buffer, or any combination thereof, for the amplification processes described above. A kit may further include instructions pertinent for the particular embodiment of the kit, such instructions describing the primer pairs and amplification conditions for operation of the method. A kit may also comprise amplification reaction containers such as microcentrifuge tubes and the like. A kit may also comprise reagents or other materials for isolating bioagent nucleic acid or bioagent identifying amplicons from amplification, including, for example, detergents, solvents, or ion exchange resins which may be linked to magnetic beads. A kit may also comprise a table of measured or calculated molecular masses and/or base compositions of bioagents using the primer pairs of the kit.

[0188] While the present invention has been described with specificity in accordance with certain of its embodiments, the following examples serve only to illustrate the invention and are not intended to limit the same. In order that the invention disclosed herein may be more efficiently understood, examples are provided below. It should be understood that these examples are for illustrative purposes only and are not to be construed as limiting the invention in any manner.

#### EXAMPLES

##### **Example 1: Selection of Primers that Define Bioagent Identifying Amplicons for VHF Viruses**

[0189] For design of primers that define viral hemorrhagic fever virus bioagent identifying amplicons, relevant sequences from, for example, GenBank were obtained, aligned and scanned for regions where pairs of PCR primers would amplify products of about 45 to about 200 nucleotides in length and distinguish species and/or sub-species from each other by their molecular masses or base compositions. A typical process shown in Figure 1 is employed.

[0190] A database of expected base compositions for each primer region is generated using an *in silico* PCR search algorithm, such as (ePCR). An existing RNA structure search algorithm (Macke et al., Nucl. Acids Res., 2001, 29, 4724-4735)

has been modified to include PCR parameters such as hybridization conditions, mismatches, and thermodynamic calculations (SantaLucia, Proc. Natl. Acad. Sci. U.S.A., 1998, 95, 1460-1465). This also provides information on primer specificity of the selected primer pairs.

[0191] Tables 4-7 represent collections of primers (sorted by forward primer name) designed to identify, flaviviruses (Table 4), filoviruses (Table 5), bunyaviruses (Table 6) and arenaviruses (Table 7) using the methods described herein. Primer sites were identified on essential filoviral, flaviviral, hantaviral, phleboviral, nairoviral and arenaviral genes, such as, for example, RNA-dependent RNA polymerase and nucleocapsid genes. The forward or reverse primer name shown in Tables 4-7 indicates the gene region of the viral genome to which the primer hybridizes relative to a reference sequence. In Table 4, for example, the forward primer name FLAV\_NC\_001474\_10032\_10056\_F indicates that the forward primer hybridizes to residues 10032-10056 of a flavivirus reference sequence represented by GenBank Accession No. NC\_001474 (SEQ ID NO: 1). In Tables 4-7, T<sup>a</sup> = 5-propynyluracil; C<sup>a</sup> = 5-propynylcytosine; I = inosine. The primer pair number is an in-house database index number.

**Table 4: Primer Pairs for Identification of Flaviviruses**

| Primer Pair Number | For. Primer Name             | For. Primer Sequence                   | For. SEQ ID NO: | Rev. Primer Name             | Rev. Primer Sequence             | Rev. SEQ ID NO: |
|--------------------|------------------------------|----------------------------------------|-----------------|------------------------------|----------------------------------|-----------------|
| 2194               | FLAV_NC_001474_10032_10056_F | TGCAGAGTGGGCC<br>AAGAACATCTGG          | 2               | FLAV_NC_001474_10138_10159_R | TGCTCTCCAGTTTGA<br>GCTCCCAGTG    | 62              |
| 2243               | FLAV_NC_001474_2084_2108_F   | TAGGAGACACAGC<br>TTGGGACTTTGG          | 3               | FLAV_NC_001474_2185_2208_R   | TTGCGTGATCCAGGA<br>CATTCTCTCC    | 63              |
| 2242               | FLAV_NC_001474_2389_2411_F   | TGGAGAGAGGTCT<br>CAGAGTGGTA            | 4               | FLAV_NC_001474_2518_2536_R   | TGCTCTCCACATGG<br>CCAT           | 64              |
| 2241               | FLAV_NC_001474_2658_2679_F   | TGAAATTGGCTGG<br>AAGGCCTGG             | 5               | FLAV_NC_001474_2731_2754_R   | TCATTCTTGGTCTC<br>CGGTCCATC      | 65              |
| 2240               | FLAV_NC_001474_2805_2829_F   | TGGAATGTTTCACG<br>ACCAACATATGG         | 6               | FLAV_NC_001474_2914_2937_R   | TCTCTCTATCCAGTA<br>ACCCATGTC     | 66              |
| 2239               | FLAV_NC_001474_2860_2879_F   | TGTGACACAGGAG<br>TCATGGG               | 7               | FLAV_NC_001474_2908_2932_R   | TCATCCAGAGACTCT<br>GATCTGTGTG    | 67              |
| 2238               | FLAV_NC_001474_2910_2933_F   | TGCTGACATGGGT<br>TACTGGATAGA           | 8               | FLAV_NC_001474_3019_3038_R   | TCCAGCACTCCATTG<br>CTCCA         | 68              |
| 2237               | FLAV_NC_001474_3214_3239_F   | TCTGTGAGGAGCA<br>CCACAGAGAGTGG         | 9               | FLAV_NC_001474_3313_3339_R   | TACCGGCCTTATTTTC<br>CATGGCATACCA | 69              |
| 2246               | FLAV_NC_001474_33_56_F       | TGCTGTCAATATG<br>CTGAAACGCGG           | 10              | FLAV_NC_001474_143_166_R     | TCCTGAAGAACGCGA<br>AAAGAGCCA     | 70              |
| 2236               | FLAV_NC_001474_3306_3335_F   | TGGCTGCTGGTAT<br>GGAATGGAGATTA<br>GACC | 11              | FLAV_NC_001474_3412_3436_R   | TGGCCAGGAACATGA<br>CCAGAAGGCC    | 71              |
| 2235               | FLAV_NC_001474_4561_4583_F   | TTCCACACTCTAT<br>GGCACACAAC            | 12              | FLAV_NC_001474_4627_4649_R   | TCCTCTTTCACACTG<br>CCCCAGTA      | 72              |
| 2234               | FLAV_NC_001474_5270_5297_F   | TCATGGATGAAGC<br>ACATTTTCACAGAT<br>CC  | 13              | FLAV_NC_001474_5341_5365_R   | TGAAGATCGCAGCTG<br>CCTCTCCCAT    | 73              |
| 2233               | FLAV_NC_001474_5273_5300_F   | TGGATGAAGCTCA<br>TTTCACCGATCCA<br>GC   | 14              | FLAV_NC_001474_5365_5385_R   | TCCCGGCGGGGTGGC<br>TGTCAT        | 74              |
| 2232               | FLAV_NC_001474_5287_5309_F   | TGGACTGATCCCC<br>ACAGCATAGC            | 15              | FLAV_NC_001474_5401_5421_R   | TATGGCTCCGTTGGA<br>CTCCGG        | 75              |
| 2231               | FLAV_NC_001474_5514_5540_F   | TAGCGTGAAAATG<br>GGGAATGAGATTG<br>C    | 16              | FLAV_NC_001474_5629_5653_R   | TGTCAGTTGTGATGA<br>CAAAGTCCCA    | 76              |
| 2230               | FLAV_NC_001474_6050_6071_F   | TCACACCGTGGCT<br>GGCATGGCA             | 17              | FLAV_NC_001474_6109_6131_R   | TCCTCTGGGCCTTCC<br>CATGTCCA      | 77              |
| 2229               | FLAV_NC_001474_6947_6968_F   | TGGGTCTTGGCAA<br>AGGATGGCC             | 18              | FLAV_NC_001474_7069_7092_R   | TCCTGGGCCTATTAT<br>GGCATAATG     | 78              |
| 2228               | FLAV_NC_001474_7271_7295_F   | TGAGGACAACATG<br>GGCCTTGTGTGA          | 19              | FLAV_NC_001474_7357_7381_R   | TGGACACGGCTATGG<br>TGGTGTCCA     | 79              |
| 2245               | FLAV_NC_001474_749_773_F     | TGGCGGCTGTTCT<br>TGGTTGGATGCT          | 20              | FLAV_NC_001474_835_862_R     | TGCTCATTCACAGGC<br>AGTTAAAGCTGTA | 80              |
| 2227               | FLAV_NC_001474_7703_7727_F   | TCGGCTGTGGAAG<br>AGGAGGCTGGTC          | 21              | FLAV_NC_001474_7777_7803_R   | TGGTTCTTCATGTCC<br>TGGTCTCTCTTT  | 81              |

|      |                              |                                       |    |                              |                                 |     |
|------|------------------------------|---------------------------------------|----|------------------------------|---------------------------------|-----|
| 2226 | FLAV_NC_001474_7713_7736 F   | TAGAGGCGGCTGG<br>TCCCTACTATGC         | 22 | FLAV_NC_001474_7808_7831 R   | TGTTCCAACCGAGGC<br>TTGTTACCA    | 82  |
| 2244 | FLAV_NC_001474_835_858 F     | TACAGCTTCAACT<br>GTCTGGGAATG          | 23 | FLAV_NC_001474_889_910 R     | TCAAATCCACCCAAG<br>TGGCTCC      | 83  |
| 2225 | FLAV_NC_001474_8358_8382 F   | TCCATACAGGACA<br>TGGCAGTACTGG         | 24 | FLAV_NC_001474_8452_8474 R   | TCTTCCCGTGCATTC<br>CATGGCCA     | 84  |
| 2224 | FLAV_NC_001474_8377_8400 F   | TACCACGGAAGTT<br>ATGAGGTGAAG          | 25 | FLAV_NC_001474_8479_8498 R   | TCAGTCATGGCCATG<br>GTGGT        | 85  |
| 2223 | FLAV_NC_001474_8528_8555 F   | TCAAAGAGAAGGT<br>TGACACGAAAGCT<br>CC  | 26 | FLAV_NC_001474_8593_8615 R   | TAGGCCCAACAACCAG<br>TTGGTGGT    | 86  |
| 2222 | FLAV_NC_001474_8803_8831 F   | TGTCACACGTGTG<br>TCTACAACATGAT<br>GGG | 27 | FLAV_NC_001474_8887_8907 R   | TGCTCCCAGCCACAT<br>GTACCA       | 87  |
| 2221 | FLAV_NC_001474_8807_8831 2 F | TCACCTGCATCTA<br>CAACATGATGGG         | 28 | FLAV_NC_001474_8926_8951 R   | TGGTCTTCATTGAGG<br>AATCCCAGAGC  | 88  |
| 2220 | FLAV_NC_001474_8807_8831 F   | TCCATTGCGTGTA<br>CAACATGATGGG         | 29 | FLAV_NC_001474_8887_8908 R   | TACTCCCCAGCCACA<br>TGTACCA      | 89  |
| 2219 | FLAV_NC_001474_8808_8831 F   | TACTTGCGTCTAC<br>AACATGATGGG          | 30 | FLAV_NC_001474_8887_8907 R   | TGCTCCCAGCCACAT<br>GTACCA       | 87  |
| 2218 | FLAV_NC_001474_8812_8840 2 F | TGTGTGTACAACA<br>TGATGGGGAAGAG<br>AGA | 31 | FLAV_NC_001474_8881_8905 R   | TGCCCAGCCACATGT<br>ACCAGATGGC   | 90  |
| 2217 | FLAV_NC_001474_8812_8840 F   | TGTGTCTACAACA<br>TGATGGGGAAGAG<br>AGA | 32 | FLAV_NC_001474_8887_8907 R   | TGCTCCCAGCCACAT<br>GTACCA       | 87  |
| 2216 | FLAV_NC_001474_8865_8885 F   | TGCCAAGGGAAGC<br>AGGGCCAT             | 33 | FLAV_NC_001474_8926_8951 R   | TGGTCTTCATTGAGG<br>AATCCCAGAGC  | 88  |
| 2215 | FLAV_NC_001474_8874_8898 F   | TAGCCGAGCCATC<br>TGGTACATGTGG         | 34 | FLAV_NC_001474_8941_8966 R   | TCTCTGGAAAGCCAG<br>TGGTCTTCATT  | 91  |
| 2213 | FLAV_NC_001474_8880_8903 F   | TGCCATCTGGTAC<br>ATGTGGCTGGG          | 35 | FLAV_NC_001474_8971_8992 R   | TTCCCTCAACTCCAG<br>CTCCACT      | 92  |
| 2214 | FLAV_NC_001474_8880_8903 F   | TGCCATCTGGTAC<br>ATGTGGCTGGG          | 35 | FLAV_NC_001474_8941_8966 R   | TCTCTGGAAAGCCAG<br>TGGTCTTCATT  | 91  |
| 2211 | FLAV_NC_001474_8885_8906 2 F | TCTGGTTCATGTG<br>GCTGGGAGC            | 36 | FLAV_NC_001474_8941_8964 R   | TCTGCCCAGCCAGTG<br>GTCTTCATT    | 93  |
| 2210 | FLAV_NC_001474_8885_8906 F   | TCTGGTACATGTG<br>GCTGGGAGC            | 37 | FLAV_NC_001474_8977_9002 R   | TGCAGACCTTCTCCT<br>TCCACTCCACT  | 94  |
| 2212 | FLAV_NC_001474_8885_8906 F   | TCTGGTACATGTG<br>GCTGGGAGC            | 37 | FLAV_NC_001474_8941_8964 R   | TCTGCCCAGCCAGTG<br>GTCTTCATT    | 93  |
| 2209 | FLAV_NC_001474_8930_8955 F   | TGGGATTTCCTGAA<br>TGAAGACCACTGG       | 38 | FLAV_NC_001474_9061_9084 R   | TGTGTCCCAGCCGGC<br>TGTGTCATC    | 95  |
| 2208 | FLAV_NC_001474_8971_8996 F   | TCATTGAGTGGAG<br>TGGAAGGAGAAGG        | 39 | FLAV_NC_001474_9061_9080 2 R | TCCCAGCCGGCTGTG<br>TCATC        | 96  |
| 2207 | FLAV_NC_001474_8984_9011 2 F | TGGAAGGCATTGG<br>CTTACAATACCTA<br>GG  | 40 | FLAV_NC_001474_9061_9080 R   | TCCCATCCAGCGGTG<br>TCATC        | 97  |
| 2206 | FLAV_NC_001474_8984_9011 F   | TGGAGGGAATCAG<br>CCTGAACTACCTG<br>GG  | 41 | FLAV_NC_001474_9064_9085 R   | TCGTGTCCCAGCCAG<br>CTGTGTC      | 98  |
| 2205 | FLAV_NC_001474_8999_9026 F   | TCCAGAAGCTGGG<br>ATACATCCTGCGT<br>GA  | 42 | FLAV_NC_001474_9112_9135 R   | TAGCAACTCCAGCAC<br>CTTAGCTTC    | 99  |
| 2204 | FLAV_NC_001474_9239_9263 F   | TCATAAGTCGACG<br>AGACCAGAGAGG         | 43 | FLAV_NC_001474_9313_9339 R   | TGCTTCTGCCATTCT<br>GATCAATTGGAC | 100 |
| 2203 | FLAV_NC_001474_9259_9281 F   | TGAGGCTCAGGTC<br>AGGTTGTGAC           | 44 | FLAV_NC_001474_9328_9351 R   | TATGACACCCTCCCC<br>CTCCATCAT    | 101 |
| 2202 | FLAV_NC_001474_9437_9458 F   | TGGCGGTGAGTGG<br>AGACGACTG            | 45 | FLAV_NC_001474_9501_9526 R   | TAACCTTGACATGG<br>CGTTGAGATGG   | 102 |
| 2201 | FLAV_NC_001474_9439_9461 F   | TCTGTCAGCGGAG<br>ATGACTGTGT           | 46 | FLAV_NC_001474_9514_9539 R   | TGGATGTCTTTTCGG<br>ACCTTTGACAT  | 103 |
| 2200 | FLAV_NC_001474_9492_9518 F   | TGCCCTTTACTTC<br>CTGAATGACATGG<br>C   | 47 | FLAV_NC_001474_9571_9596 R   | TGTGAGCAGAAGGGG<br>ACCTCTTCCCA  | 104 |
| 2199 | FLAV_NC_001474_9729_9752 F   | TAAGGCCTACGGA<br>CAGATGTGGCT          | 48 | FLAV_NC_001474_9847_9870 R   | TGCACCACTGGCATG<br>GATGCTCCA    | 105 |
| 2198 | FLAV_NC_001474_9746_9767 F   | TGTGGCTGCTGCT<br>GTACTTCCA            | 49 | FLAV_NC_001474_9864_9889 R   | TGTCTTCTGTTGTCA<br>TCCACTCTCCT  | 106 |
| 2197 | FLAV_NC_001474_9827_9851 F   | TCCCAACAAGCCG<br>AACAACCTGGTC         | 50 | FLAV_NC_001474_9871_9894 R   | TAGCATGTCTTCCGT<br>GGTCATCCA    | 107 |
| 2196 | FLAV_NC_001474_9871_9891 F   | TGGATGACGACGG<br>AAGACATG             | 51 | FLAV_NC_001474_9901_9923 R   | TCCTCAATCCAGACC<br>CTGTTCCA     | 108 |
| 2195 | FLAV_NC_001474_9909_9936 F   | TGTCTGGATTGAG<br>GAGAATGAATGGA<br>TG  | 52 | FLAV_NC_001474_9994_10015 R  | TGAGGCTTCCACACC<br>AGATGTC      | 109 |

|     |                                  |                                                                                                 |    |                                  |                                                                                                        |     |
|-----|----------------------------------|-------------------------------------------------------------------------------------------------|----|----------------------------------|--------------------------------------------------------------------------------------------------------|-----|
| 526 | FLAV_NC001474_5<br>201 5216P F   | TGAT <sup>a</sup> GT <sup>a</sup> GTC <sup>a</sup> AT<br>GC <sup>a</sup> CAC                    | 53 | FLAV_NC001474_5<br>272 5288P R   | TAATGGGCTT <sup>a</sup> C <sup>a</sup> AT <sup>a</sup><br>C <sup>a</sup> C <sup>a</sup> AT             | 110 |
| 524 | FLAV_NC001474_8<br>812 8831P F   | TGTATGT <sup>a</sup> ACAAC<br>AT <sup>a</sup> GAT <sup>a</sup> GGG                              | 54 | FLAV_NC001474_8<br>887 8904P R   | TCCCAGCCACATGTA<br>C <sup>a</sup> C <sup>a</sup> A                                                     | 111 |
| 525 | FLAV_NC001474_8<br>818 8831P F   | T <sup>a</sup> ACAACAT <sup>a</sup> GAT <sup>a</sup><br>GGG                                     | 55 | FLAV_NC001474_8<br>887 8901P R   | TAGCCACATGTAC <sup>a</sup> C <sup>a</sup><br>A                                                         | 112 |
| 523 | FLAV_NC001474_8<br>818 8840 2P F | T <sup>a</sup> ACAAC <sup>a</sup> AT <sup>a</sup> GAT <sup>a</sup><br><sup>a</sup> GGGGAAGIGIGA | 56 | FLAV_NC001474_8<br>887 8904P R   | TCCCAGCCACATGTA<br>C <sup>a</sup> C <sup>a</sup> A                                                     | 111 |
| 522 | FLAV_NC001474_8<br>818 8840P F   | T <sup>a</sup> ACAAC <sup>a</sup> AT <sup>a</sup> GAT <sup>a</sup><br><sup>a</sup> GGGAAAGAGAGA | 57 | FLAV_NC001474_8<br>887 8904P R   | TCCCAGCCACATGTA<br>C <sup>a</sup> C <sup>a</sup> A                                                     | 111 |
| 569 | FLAV_NC001474_8<br>887 8903P F   | TGGTACAT <sup>a</sup> GT <sup>a</sup> GG<br>C <sup>a</sup> TGGG                                 | 58 | FLAV_NC001474_8<br>941 8956P R   | TC <sup>a</sup> C <sup>a</sup> AGTGGTCTT <sup>a</sup> C <sup>a</sup><br><sup>a</sup> AT <sup>a</sup> T | 113 |
| 528 | FLAV_NC001474_8<br>939 8955 2P F | T <sup>a</sup> GAATGAAGATC <sup>a</sup><br>ACT <sup>a</sup> GG                                  | 59 | FLAV_NC001474_9<br>061 9080P R   | TC <sup>a</sup> C <sup>a</sup> C <sup>a</sup> AICCGICI G<br>TGTCITC                                    | 114 |
| 530 | FLAV_NC001474_8<br>939 8955 2P F | T <sup>a</sup> GAATGAAGATC <sup>a</sup><br>ACT <sup>a</sup> GG                                  | 59 | FLAV_NC001474_9<br>061 9080P R   | TC <sup>a</sup> C <sup>a</sup> C <sup>a</sup> AICCGICI G<br>TGTCITC                                    | 114 |
| 566 | FLAV_NC001474_8<br>939 8955 3P F | T <sup>a</sup> GAATGAAGATC <sup>a</sup><br>ACT <sup>a</sup> GG                                  | 59 | FLAV_NC001474_9<br>061 9080 2P R | TC <sup>a</sup> C <sup>a</sup> C <sup>a</sup> AICCGICI G<br>TGTCATC                                    | 115 |
| 567 | FLAV_NC001474_8<br>939 8955 4P F | T <sup>a</sup> GAATGAAGACC <sup>a</sup><br>ATT <sup>a</sup> GG                                  | 60 | FLAV_NC001474_9<br>061 9080 3 R  | TCCCAICCGICIGTG<br>TCATC                                                                               | 116 |
| 529 | FLAV_NC001474_8<br>939 8955P F   | T <sup>a</sup> GAATGAAGACC <sup>a</sup><br>ATT <sup>a</sup> GG                                  | 60 | FLAV_NC001474_9<br>061 9080 R    | TCCCAICCGICIGTG<br>TCITC                                                                               | 117 |
| 527 | FLAV_NC001474_8<br>939 8955P F   | T <sup>a</sup> GAATGAAGACC <sup>a</sup><br>ATT <sup>a</sup> GG                                  | 60 | FLAV_NC001474_9<br>061 9080 R    | TCCCAICCGICIGTG<br>TCITC                                                                               | 117 |
| 568 | FLAV_NC001474_9<br>736 9750P F   | T <sup>a</sup> ACGCGCAGAT <sup>a</sup> G<br>T <sup>a</sup> GG                                   | 61 | FLAV_NC001474_9<br>871 9889P R   | TGTCTTCTGTTGTCA<br>T <sup>a</sup> C <sup>a</sup> C <sup>a</sup> A                                      | 118 |

[0192] Reference Sequence NC\_001474 (SEQ ID NO: 1) represents the genome of the Dengue virus.

**Table 5: Primer Pairs for Identification of Filoviruses**

| Primer pair number | For. primer name              | For. sequence                                                                                                                | For. SEQ ID NO: | Rev. primer name                | Rev. sequence                                                                                                                                                           | Rev. SEQ ID NO: |
|--------------------|-------------------------------|------------------------------------------------------------------------------------------------------------------------------|-----------------|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| 504                | FILO_NC002549_1051_1072P F    | TGGAC <sup>a</sup> AC <sup>a</sup> AU <sup>a</sup> GATG<br>GT <sup>a</sup> AAT <sup>a</sup> TT <sup>a</sup> T <sup>a</sup> C | 119             | FILO_NC002549_1_131_1151 2P R   | TGGCAT <sup>a</sup> C <sup>a</sup> ATGACCA<br>GCCAC <sup>a</sup> C <sup>a</sup> A                                                                                       | 152             |
| 503                | FILO_NC002549_1051_1072P F    | TGGAC <sup>a</sup> AC <sup>a</sup> AT <sup>a</sup> GATG<br>GT <sup>a</sup> AAT <sup>a</sup> TT <sup>a</sup> T <sup>a</sup> C | 120             | FILO_NC002549_1_131_1151P R     | TGGCAT <sup>a</sup> C <sup>a</sup> ATGGCCG<br>GCCAC <sup>a</sup> C <sup>a</sup> A                                                                                       | 153             |
| 747                | FILO_NC002549_13309_13331_F   | TGTGAAGCTCTGTTA<br>GCAGATGG                                                                                                  | 121             | FILO_NC002549_1_3399_13418 R    | TGGTGCCATGATGCC<br>TGATG                                                                                                                                                | 154             |
| 508                | FILO_NC002549_13311_13331P F  | TGAAGCTC <sup>a</sup> T <sup>a</sup> GTT <sup>a</sup> A<br>GC <sup>a</sup> AGAT <sup>a</sup> GG                              | 122             | FILO_NC002549_1_3351_13374 2P R | TT <sup>a</sup> C <sup>a</sup> AGT <sup>a</sup> GAC <sup>a</sup> TAC <sup>a</sup><br>C <sup>a</sup> AT <sup>a</sup> C <sup>a</sup> ATAT <sup>a</sup> T <sup>a</sup> GCT | 155             |
| 825                | FILO_NC002549_13311_13331P F  | TGAAGCTC <sup>a</sup> T <sup>a</sup> GTT <sup>a</sup> A<br>GC <sup>a</sup> AGAT <sup>a</sup> GG                              | 122             | FILO_NC002549_1_3351_13374 2P R | TT <sup>a</sup> C <sup>a</sup> AGT <sup>a</sup> GAC <sup>a</sup> TAC <sup>a</sup><br>C <sup>a</sup> AT <sup>a</sup> C <sup>a</sup> ATAT <sup>a</sup> T <sup>a</sup> GCT | 155             |
| 825                | FILO_NC002549_13311_13331P F  | TGAAGCTC <sup>a</sup> T <sup>a</sup> GTT <sup>a</sup> A<br>GC <sup>a</sup> AGAT <sup>a</sup> GG                              | 122             | FILO_NC002549_1_3351_13374P R   | TT <sup>a</sup> C <sup>a</sup> AGT <sup>a</sup> GAC <sup>a</sup> TAC <sup>a</sup><br>C <sup>a</sup> AT <sup>a</sup> C <sup>a</sup> ATGT <sup>a</sup> T <sup>a</sup> ACT | 156             |
| 507                | FILO_NC002549_13311_13331P F  | TGAAGCTC <sup>a</sup> T <sup>a</sup> GTT <sup>a</sup> A<br>GC <sup>a</sup> AGAT <sup>a</sup> GG                              | 122             | FILO_NC002549_1_3351_13374P R   | TT <sup>a</sup> C <sup>a</sup> AGT <sup>a</sup> GAC <sup>a</sup> TAC <sup>a</sup><br>C <sup>a</sup> AT <sup>a</sup> C <sup>a</sup> ATGT <sup>a</sup> T <sup>a</sup> ACT | 156             |
| 509                | FILO_NC002549_13311_13331P F  | TGAAGCTC <sup>a</sup> T <sup>a</sup> GTT <sup>a</sup> A<br>GC <sup>a</sup> AGAT <sup>a</sup> GG                              | 122             | FILO_NC002549_1_3357_13377P R   | TCGTT <sup>a</sup> C <sup>a</sup> AGT <sup>a</sup> GAC <sup>a</sup><br>TAC <sup>a</sup> C <sup>a</sup> AT <sup>a</sup> C <sup>a</sup> AT                                | 157             |
| 859                | FILO_NC002549_13397_13418_2 F | TGCATCAGGCATCTT<br>GGCACCA                                                                                                   | 123             | FILO_NC002549_1_3465_13498 R    | TGGCGAGATTGTATT<br>TCTCTAGATCAGTGA<br>CAAA                                                                                                                              | 158             |
| 858                | FILO_NC002549_13397_13418_F   | TTCATCAGGCATCAT<br>GGCACCA                                                                                                   | 124             | FILO_NC002549_1_3471_13499 R    | TCGGCGAGGTTGTATT<br>TTCTCTAGATCAGT                                                                                                                                      | 159             |
| 746                | FILO_NC002549_13401_13420_F   | TCAGGCATCATGGCA<br>CCACA                                                                                                     | 125             | FILO_NC002549_1_3471_13497 R    | TGCAAGGTTGTATTT<br>CTCTAGATCAGT                                                                                                                                         | 160             |
| 510                | FILO_NC002549_13401_13420P F  | TC <sup>a</sup> AGGCTT <sup>a</sup> C <sup>a</sup> ATGG<br>CAC <sup>a</sup> C <sup>a</sup> ACA                               | 126             | FILO_NC002549_1_3471_13494P R   | TAGAT <sup>a</sup> T <sup>a</sup> GT <sup>a</sup> ATT <sup>a</sup> T <sup>a</sup><br>C <sup>a</sup> T <sup>a</sup> C <sup>a</sup> TAGAT <sup>a</sup> C <sup>a</sup> AGT | 161             |
| 857                | FILO_NC002549_13461_13490_F   | TAGCCCTGTCACTGA<br>TCTAGAGAAATACAA                                                                                           | 127             | FILO_NC002549_1_3594_13621 3 R  | TGCGGTAATCACTGA<br>CATGCATATAACA                                                                                                                                        | 162             |

|     |                                        |                                                                                                                                                         |     |                                    |                                                                                                                                                                      |     |
|-----|----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|-----|------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 852 | FILO_NC002549<br>_13546_13577_<br>F    | TATGGTGTGAGGAAT<br>GTCTTTGATTGGATG<br>CA                                                                                                                | 128 | FILO_NC002549_1<br>3594_13621_2_R  | TGCAAAAATCACTGA<br>CATGCATGTAACA                                                                                                                                     | 163 |
| 853 | FILO_NC002549<br>_13546_13577P<br>F    | TAT <sup>a</sup> GGT <sup>a</sup> GT <sup>a</sup> IIIIA<br>ATGTCTTTGATTGGA<br>TGCA                                                                      | 129 | FILO_NC002549_1<br>3594_13624P_R   | TGC <sup>a</sup> T <sup>a</sup> AT <sup>a</sup> AAIIITC<br>ACTGACATGCATGTA<br>ACA                                                                                    | 164 |
| 850 | FILO_NC002549<br>_13551_13577_<br>F    | TGTGCGGAATGTCTT<br>TGATTGGATGCA                                                                                                                         | 130 | FILO_NC002549_1<br>3594_13621_R    | TATGCCAATCACTGA<br>CATGCATGTAACA                                                                                                                                     | 165 |
| 851 | FILO_NC002549<br>_13551_13577P<br>F    | TGTGCGGAATGTCT <sup>a</sup><br>T <sup>a</sup> TGAT <sup>a</sup> T <sup>a</sup> GGAT <sup>a</sup> GC<br>A                                                | 130 | FILO_NC002549_1<br>3594_13621P_R   | TATGCCAAT <sup>a</sup> C <sup>a</sup> AC <sup>a</sup> T <sup>a</sup><br><sup>a</sup> GAC <sup>a</sup> AT <sup>a</sup> GC <sup>a</sup> ATGTA<br>ACA                   | 165 |
| 511 | FILO_NC002549<br>_13557_13579P<br>F    | TAAT <sup>a</sup> GTCT <sup>a</sup> T <sup>a</sup> T <sup>a</sup> GA<br>T <sup>a</sup> T <sup>a</sup> GGAT <sup>a</sup> GC <sup>a</sup> ATT             | 131 | FILO_NC002549_1<br>3594_13614P_R   | TT <sup>a</sup> C <sup>a</sup> AC <sup>a</sup> T <sup>a</sup> GAC <sup>a</sup> AT <sup>a</sup><br>GC <sup>a</sup> AT <sup>a</sup> AT <sup>a</sup> AAC <sup>a</sup> A | 166 |
| 855 | FILO_NC002549<br>_13591_13621_<br>F    | TIICGTTACATGCAT<br>GTCAGTGACTATTAT<br>A                                                                                                                 | 132 | FILO_NC002549_1<br>3726_13750_R    | TTTGIGCACAGGAIA<br>TGCTTGTTCCA                                                                                                                                       | 167 |
| 854 | FILO_NC002549<br>_13591_13621P<br>F    | TIICGT <sup>a</sup> T <sup>a</sup> ACAT <sup>a</sup> GC <sup>a</sup><br><sup>a</sup> AT <sup>a</sup> GTCAGTGACTAT<br>TATA                               | 132 | FILO_NC002549_1<br>3726_13750P_R   | TTT <sup>a</sup> GIGC <sup>a</sup> AC <sup>a</sup> AGGA<br>IATGCTTGTTCCA                                                                                             | 167 |
| 565 | FILO_NC002549<br>_13594_13613P<br>F    | TGT <sup>a</sup> T <sup>a</sup> ACAT <sup>a</sup> GC <sup>a</sup> AT <sup>a</sup><br>GTCAGTGA                                                           | 133 | FILO_NC002549_1<br>3696_13715P_R   | TGTAGTCCCTCTATC<br>CC <sup>a</sup> T <sup>a</sup> C <sup>a</sup> C                                                                                                   | 168 |
| 856 | FILO_NC002549<br>_13594_13621_<br>F    | TGCCGCATGCATGTC<br>AGTGATTATTATA                                                                                                                        | 134 | FILO_NC002549_1<br>3726_13750_2_R  | TTTGAGCACAGGATA<br>TGCTTGTTCCA                                                                                                                                       | 169 |
| 866 | FILO_NC002549<br>_13722_13745_<br>2_F  | TCTGTGGACAAGTAT<br>ATCATGTGC                                                                                                                            | 135 | FILO_NC002549_1<br>3795_13816_2_R  | TACACTGATTGTCAC<br>CCATGAC                                                                                                                                           | 170 |
| 745 | FILO_NC002549<br>_13722_13745_<br>F    | TCTGTGGACAAGTAT<br>ATCATGTGC                                                                                                                            | 135 | FILO_NC002549_1<br>3795_13816_R    | TACACTGATTGTCAC<br>CCATCAC                                                                                                                                           | 171 |
| 865 | FILO_NC002549<br>_13722_13745P<br>F    | TCTGTGGAC <sup>a</sup> AAGTA<br>T <sup>a</sup> AT <sup>a</sup> C <sup>a</sup> AT <sup>a</sup> GT <sup>a</sup> GC                                        | 135 | FILO_NC002549_1<br>3795_13816P_R   | TACACTGAT <sup>a</sup> T <sup>a</sup> GT <sup>a</sup> C <sup>a</sup><br><sup>a</sup> AC <sup>a</sup> C <sup>a</sup> C <sup>a</sup> ATGAC                             | 170 |
| 861 | FILO_NC002549<br>_13722_13751_<br>2P_F | TCCGTGGAC <sup>a</sup> AAGTA<br>T <sup>a</sup> AT <sup>a</sup> C <sup>a</sup> AT <sup>a</sup> GT <sup>a</sup> GCTC<br>AAAT                              | 136 | FILO_NC002549_1<br>3795_13823_2P_R | TCTGTGATACACTGA<br>T <sup>a</sup> T <sup>a</sup> GT <sup>a</sup> C <sup>a</sup> AC <sup>a</sup> C <sup>a</sup> A<br>TGAC                                             | 172 |
| 862 | FILO_NC002549<br>_13722_13751_<br>F    | TCCGTGGACAAGTAT<br>ATCATGTGCTCAAAAT                                                                                                                     | 136 | FILO_NC002549_1<br>3795_13823_R    | TCTGTGATACACTGA<br>TTGTCACCCATGAC                                                                                                                                    | 172 |
| 860 | FILO_NC002549<br>_13722_13751P<br>F    | TCCGT <sup>a</sup> GGACAAGTA<br>TAT <sup>a</sup> CATGT <sup>a</sup> GCTCAA<br>AT                                                                        | 136 | FILO_NC002549_1<br>3795_13823P_R   | TCTGTGATACACTGA<br>T <sup>a</sup> TGTCACC <sup>a</sup> C <sup>a</sup> ATGA<br>C                                                                                      | 172 |
| 512 | FILO_NC002549<br>_13726_13745P<br>F    | TGGAC <sup>a</sup> AAGTAT <sup>a</sup> AT <sup>a</sup><br>C <sup>a</sup> AT <sup>a</sup> GT <sup>a</sup> GC                                             | 137 | FILO_NC002549_1<br>3799_13817P_R   | TTACACTGAT <sup>a</sup> T <sup>a</sup> GT <sup>a</sup><br>C <sup>a</sup> AC <sup>a</sup> C <sup>a</sup> C <sup>a</sup> A                                             | 173 |
| 864 | FILO_NC002549<br>_13726_13751_<br>F    | TGGACAAGTATATCA<br>TGTGCTCAAAAT                                                                                                                         | 138 | FILO_NC002549_1<br>3798_13823_R    | TCTGTGATACACTGA<br>TTGTCACCCAT                                                                                                                                       | 174 |
| 863 | FILO_NC002549<br>_13726_13751P<br>F    | TGGAC <sup>a</sup> AAGTAT <sup>a</sup> AT <sup>a</sup><br>C <sup>a</sup> AT <sup>a</sup> GT <sup>a</sup> GCTCAAAAT                                      | 138 | FILO_NC002549_1<br>3798_13823P_R   | TCTGTGATACACTGA<br>T <sup>a</sup> T <sup>a</sup> GT <sup>a</sup> C <sup>a</sup> AC <sup>a</sup> C <sup>a</sup> A<br>T                                                | 174 |
| 506 | FILO_NC002549<br>_1432_1449P_F         | TGGAGTGGC <sup>a</sup> CAC <sup>a</sup> AG<br>C <sup>a</sup> ACA                                                                                        | 139 | FILO_NC002549_1<br>466_1487P_R     | TCT <sup>a</sup> GTT <sup>a</sup> C <sup>a</sup> TC <sup>a</sup> C <sup>a</sup> AA<br>C <sup>a</sup> AT <sup>a</sup> T <sup>a</sup> GACTCC                           | 175 |
| 505 | FILO_NC002549<br>_1432_1449P_F         | TGGAGTGGC <sup>a</sup> CAC <sup>a</sup> AG<br>C <sup>a</sup> ACA                                                                                        | 139 | FILO_NC002549_1<br>472_1492P_R     | TTGATACT <sup>a</sup> GTT <sup>a</sup> C <sup>a</sup> T <sup>a</sup><br>C <sup>a</sup> C <sup>a</sup> AAC <sup>a</sup> ATT                                           | 176 |
| 514 | FILO_NC002549<br>_14644_14666P<br>F    | TC <sup>a</sup> AT <sup>a</sup> C <sup>a</sup> AAATC <sup>a</sup> C <sup>a</sup> T <sup>a</sup><br><sup>a</sup> GT <sup>a</sup> TAT <sup>a</sup> GAGTCG | 140 | FILO_NC002549_1<br>4720_14736P_R   | TCC <sup>a</sup> T <sup>a</sup> T <sup>a</sup> C <sup>a</sup> AAGGTA<br>TC <sup>a</sup> C <sup>a</sup> T <sup>a</sup> A                                              | 177 |
| 513 | FILO_NC002549<br>_14652_14672P<br>F    | TC <sup>a</sup> C <sup>a</sup> T <sup>a</sup> GT <sup>a</sup> AAT <sup>a</sup> GAG<br>TC <sup>a</sup> GCT <sup>a</sup> T <sup>a</sup> T <sup>a</sup> GC | 141 | FILO_NC002549_1<br>4720_14739P_R   | TGTTCC <sup>a</sup> T <sup>a</sup> T <sup>a</sup> C <sup>a</sup> AAG<br>ATATC <sup>a</sup> C <sup>a</sup> T <sup>a</sup> A                                           | 178 |
| 867 | FILO_NC002549<br>_890_909_F            | GAGACAACGGAAGCT<br>AATGC                                                                                                                                | 142 | FILO_NC002549_1<br>057_1076_R      | AACGGAAGATCACCA<br>TCATG                                                                                                                                             | 179 |
| 868 | FILO_NC002549<br>_911_930_F            | GGTCAGTTTCTATCC<br>TTTGC                                                                                                                                | 143 | FILO_NC002549_1<br>041_1060_R      | CATGTGTCCAACCTGA<br>TTGCC                                                                                                                                            | 180 |
| 871 | FILO_NC002549<br>_938_963_2_F          | TTTCTACCCAAACTT<br>GTCGTTGGGGA                                                                                                                          | 144 | FILO_NC002549_1<br>055_1080_2_R    | TTCAAACGGAAGATC<br>ACCATCATGTG                                                                                                                                       | 181 |
| 870 | FILO_NC002549<br>_938_963_F            | TTCTTCCCAAACCTG<br>GTCGTTGGAGA                                                                                                                          | 145 | FILO_NC002549_1<br>052_1078_R      | TAGGCGGAAAATTAC<br>CATCATGTGTCC                                                                                                                                      | 182 |
| 869 | FILO_NC002549<br>_938_963_F            | TTCTTCCCAAACCTG<br>GTCGTTGGAGA                                                                                                                          | 145 | FILO_NC002549_1<br>055_1080_R      | TGCAACCGGAAAATT<br>ACCATCATGTG                                                                                                                                       | 183 |

|     |                               |                                                                                   |     |                                  |                                                                                   |     |
|-----|-------------------------------|-----------------------------------------------------------------------------------|-----|----------------------------------|-----------------------------------------------------------------------------------|-----|
| 873 | FILO_NC002549<br>942_963 F    | TCCCAAACTTGTCTG<br>TCGGAGA                                                        | 146 | FILO_NC002549_1<br>040_1062 R    | TTCATGTGGCCTGTG<br>GTAAGCCA                                                       | 184 |
| 501 | FILO_NC002549<br>943_963 2P F | TC <sup>a</sup> C <sup>a</sup> GAAACTGGT <sup>a</sup> C<br>GT <sup>a</sup> GGGAGA | 147 | FILO_NC002549_1<br>052_1073 2P R | TGAAAATTACCTT <sup>a</sup> C <sup>a</sup><br>AT <sup>a</sup> GT <sup>a</sup> GTCC | 185 |
| 502 | FILO_NC002549<br>943_963 3P F | TC <sup>a</sup> C <sup>a</sup> GAAACTGGT <sup>a</sup> T<br>GT <sup>a</sup> CGGAGA | 148 | FILO_NC002549_1<br>052_1073 3P R | TGAAAATTACTTT <sup>a</sup> C <sup>a</sup><br>AT <sup>a</sup> GT <sup>a</sup> GTCC | 186 |
| 748 | FILO_NC002549<br>943_963 F    | TCCGAACTGGTAGT<br>GGGAGA                                                          | 149 | FILO_NC002549_1<br>040_1061 R    | TCATGTGTCCTACTG<br>ATTGCCA                                                        | 187 |
| 500 | FILO_NC002549<br>943_963P F   | TC <sup>a</sup> C <sup>a</sup> GAAACTGGT <sup>a</sup> C<br>GT <sup>a</sup> AGGAGA | 150 | FILO_NC002549_1<br>052_1073P R   | TGAAAATTACCAT <sup>a</sup> C <sup>a</sup><br>AT <sup>a</sup> GT <sup>a</sup> GTCC | 188 |
| 872 | FILO_NC002549<br>984_1011 F   | TTCAGAGGCAAATTC<br>AGGTACATGCAGA                                                  | 151 | FILO_NC002549_1<br>040_1066 R    | TACCATCATGTGTCC<br>TACTGATTGCCA                                                   | 189 |

[0193] Reference sequence NC\_002549 (SEQ ID NO: 268) represents the genome of Ebola Zaire virus.

**Table 6: Primer Pairs for Identification of Bunyaviruses**

| Primer<br>Pair<br>Number | For. Primer<br>Name                   | For. Primer<br>Sequence                                                         | For.<br>SEQ ID<br>NO: | Rev. Primer<br>Name               | Rev. Primer<br>Sequence                                                          | Rev.<br>SEQ ID<br>NO: |
|--------------------------|---------------------------------------|---------------------------------------------------------------------------------|-----------------------|-----------------------------------|----------------------------------------------------------------------------------|-----------------------|
| 592                      | HVLGENE_X5590<br>1_1740_1760 F        | TACAGCCACATGGTT<br>CCAATA                                                       | 190                   | HVLGENE_X55901<br>1849_1871P R    | TCAAAGATTGCACAT<br>AGTTT <sup>a</sup> C <sup>a</sup> AT                          | 216                   |
| 591                      | HVLGENE_X5590<br>1_2_23P F            | T <sup>a</sup> GGAGAAAT <sup>a</sup> ATAGA<br>GAGATTCA                          | 191                   | HVLGENE_X55901<br>125_144 R       | TGACCAGTCATGCTT<br>TATCA                                                         | 217                   |
| 593                      | HVLGENE_X5590<br>1_2077_2096 F        | TCAACTGTCTGGTGCA<br>AGTGG                                                       | 192                   | HVLGENE_X55901<br>2182_2201P R    | TTCCCATGCAGACCC<br>T <sup>a</sup> T <sup>a</sup> TTC                             | 218                   |
| 594                      | HVLGENE_X5590<br>1_2820_2840 F        | TAAGGCACTCAGATG<br>GGCATC                                                       | 193                   | HVLGENE_X55901<br>2899_2920P R    | TGGCAT <sup>a</sup> C <sup>a</sup> TGCACTA<br>ACATACAT                           | 219                   |
| 373                      | HVLGENE_X5590<br>1_2897_2918 F        | TCATGTATGTTAGTG<br>CTGATGC                                                      | 194                   | HVLGENE_X55901<br>2926_2946 R     | TGAATTATCTCCTGG<br>TGACCA                                                        | 220                   |
| 374                      | HVLGENE_X5590<br>1_2897_2918 F        | TCATGTATGTTAGTG<br>CTGATGC                                                      | 194                   | HVLGENE_X55901<br>2928_2949 R     | TGCTGAATTATCTCC<br>TGGTGAC                                                       | 221                   |
| 595                      | HVLGENE_X5590<br>1_3279_3299P_<br>F   | TGCTCATCATT <sup>a</sup> C <sup>a</sup> AG<br>ATGATGC                           | 195                   | HVLGENE_X55901<br>3331_3351P R    | TAAC <sup>a</sup> C <sup>a</sup> AATCAGTTC<br><sup>a</sup> C <sup>a</sup> ATCATC | 222                   |
| 596                      | HVLGENE_X5590<br>1_3279_3299P_<br>F   | TGCTCATCATT <sup>a</sup> C <sup>a</sup> AG<br>ATGATGC                           | 195                   | HVLGENE_X55901<br>3412_3431P R    | TTAAACATGCTCT <sup>a</sup> T <sup>a</sup><br>C <sup>a</sup> C <sup>a</sup> ACAT  | 223                   |
| 370                      | HVLGENE_X5590<br>1_3329_3350 F        | TAGATGATGGAAGT<br>ACTGGTT                                                       | 196                   | HVLGENE_X55901<br>3412_3435 R     | TAGATTAAACATGCT<br>TTTCCACAT                                                     | 224                   |
| 371                      | HVLGENE_X5590<br>1_3408_3430 F        | TGAGATGTGGAAAAG<br>CATGTTTA                                                     | 196                   | HVLGENE_X55901<br>3526_3549 R     | TATTGATACAGCACA<br>ACCTTCAAA                                                     | 225                   |
| 372                      | HVLGENE_X5590<br>1_4173_4195 F        | TATTGTAACAGCTAT<br>GACCATGC                                                     | 197                   | HVLGENE_X55901<br>4224_4244 R     | TTCATGTGTTGCTTT<br>GCTTGC                                                        | 226                   |
| 597                      | HVLGENE_X5590<br>1_4182_4202 F        | TGCTATGACAATGCA<br>GTCACC                                                       | 198                   | HVLGENE_X55901<br>4267_4287 R     | TACCTCCCTGAATGT<br>TACCCA                                                        | 227                   |
| 598                      | HVLGENE_X5590<br>1_5329_5348P_<br>F   | T <sup>a</sup> GTGGGAT <sup>a</sup> GAGAT <sup>a</sup><br>T <sup>a</sup> TAAAAC | 199                   | HVLGENE_X55901<br>5401_5420P R    | TTAGGCTTTCC <sup>a</sup> C <sup>a</sup> CA<br>TTCAAA                             | 228                   |
| 375                      | HVSGENE_NC003<br>466_1050_1070<br>F   | TATGCGGAATACCAT<br>CATGGC                                                       | 200                   | HVSGENE_NC0034_6<br>6_1153_1172 R | TGGTCCAGTTGTATT<br>CCCAT                                                         | 229                   |
| 599                      | HVSGENE_NC003<br>466_1143_1163<br>F   | TACACAATCGATGGG<br>AATACA                                                       | 201                   | HVSGENE_NC0034_6<br>6_1234_1253 R | TCAGGATCCATATCA<br>TCACC                                                         | 230                   |
| 605                      | NAIRON_U88410<br>1169_1187 F          | TGGCTCTACATGCAC<br>CCTG                                                         | 202                   | NAIRON_U88410_1<br>226_1246P R    | TACAGGGATAGTCC <sup>a</sup><br>C <sup>a</sup> AAAGCA                             | 231                   |
| 601                      | NAIRON_U88410<br>12_29 F              | TACGTGCCGCTTTTCG<br>CCC                                                         | 203                   | NAIRON_U88410_1<br>46_168 R       | TCACAGAAGGAGGCG<br>GAGTTTGT                                                      | 232                   |
| 600                      | NAIRON_U88410<br>3_20P F              | TCAAAGACACACGTG<br>C <sup>a</sup> CG                                            | 204                   | NAIRON_U88410_5<br>6_76 R         | TGCCTCGATTGTGTT<br>CTCCAT                                                        | 233                   |
| 602                      | NAIRON_U88410<br>438_456 F            | TGGCTGCCCTAAAGT<br>GGAG                                                         | 205                   | NAIRON_U88410_5<br>18_537 R       | TCGCCAGGGACTTTG<br>TACTC                                                         | 234                   |
| 603                      | NAIRON_U88410<br>517_534P F           | TGAGT <sup>a</sup> ACAAAGTCC <sup>a</sup><br>C <sup>a</sup> TGG                 | 206                   | NAIRON_U88410_5<br>69_588P R      | TTCC <sup>a</sup> T <sup>a</sup> GCTCCTAAT<br>CATGTC                             | 235                   |
| 604                      | NAIRON_U88410<br>568_588P F           | TGACAT <sup>a</sup> GAT <sup>a</sup> TAGGA<br>GCAGGAA                           | 207                   | NAIRON_U88410_6<br>92_708P R      | TCC <sup>a</sup> C <sup>a</sup> AAGGAGGGTT<br>GAA                                | 236                   |
| 583                      | PHLEBOL_NC002<br>043_2769_2789<br>P F | TGGT <sup>a</sup> C <sup>a</sup> T <sup>a</sup> GAGAGAG<br>ATCTATGT             | 208                   | PHLEBOL_NC0020_4<br>3_2899_2917 R | TGCCGTGTGTTTCAG<br>GAAT                                                          | 237                   |

|     |                                       |                                                                                  |     |                                   |                                                                                              |     |
|-----|---------------------------------------|----------------------------------------------------------------------------------|-----|-----------------------------------|----------------------------------------------------------------------------------------------|-----|
| 584 | PHLEBOL_NC002<br>043_2894_2915<br>F   | TGACGATTCTGAAA<br>CACATGG                                                        | 209 | PHLEBOL_NC00204<br>3 2965 2986 R  | TCCACTTGCTAGCAT<br>CATCTGA                                                                   | 238 |
| 585 | PHLEBOL_NC002<br>043_2958_2975<br>P F | TGCGACAT <sup>a</sup> C <sup>a</sup> AGATG<br>AT <sup>a</sup> GC                 | 210 | PHLEBOL_NC00204<br>3 3058 3076P R | TCGAGCATC <sup>a</sup> C <sup>a</sup> TC <sup>a</sup> T<br><sup>a</sup> AAT <sup>a</sup> GAT | 239 |
| 590 | PHLEBOL_NC002<br>043_2965_2985<br>F   | TCAGATGATGCTAGC<br>AAGTGG                                                        | 211 | PHLEBOL_NC00204<br>3 3058 3080 R  | TACATCGAGCATCCT<br>CTAATGAT                                                                  | 240 |
| 586 | PHLEBOL_NC002<br>043_3243_3260<br>P F | TACTGGGAT <sup>a</sup> GAT <sup>a</sup> GC<br><sup>a</sup> AGGG                  | 212 | PHLEBOL_NC00204<br>3 3386 3403P R | TATCAT <sup>a</sup> C <sup>a</sup> T <sup>a</sup> GAGC <sup>a</sup><br>C <sup>a</sup> CTGCA  | 241 |
| 587 | PHLEBOL_NC002<br>043_3287_3302<br>P F | TTCACAC <sup>a</sup> C <sup>a</sup> CTGCAC<br>CA                                 | 213 | PHLEBOL_NC00204<br>3 3386 3403P R | TATCAT <sup>a</sup> C <sup>a</sup> T <sup>a</sup> GAGC <sup>a</sup><br>C <sup>a</sup> CTGCA  | 241 |
| 588 | PHLEBOL_NC002<br>043_3386_3401<br>P F | T <sup>a</sup> GC <sup>a</sup> AAGGCT <sup>a</sup> C <sup>a</sup> AG<br>ATGA     | 214 | PHLEBOL_NC00204<br>3 3509 3525P R | TGATGGGT <sup>a</sup> AAAT <sup>a</sup> GC<br>C <sup>a</sup> AA                              | 242 |
| 589 | PHLEBOL_NC002<br>043_4103_4122<br>P F | TGAACATAC <sup>a</sup> C <sup>a</sup> AGAG<br>AAC <sup>a</sup> T <sup>a</sup> GG | 215 | PHLEBOL_NC00204<br>3 4201 4221P R | TCCTGGACTATGGAC<br>CT <sup>a</sup> T <sup>a</sup> CTC                                        | 243 |

[0194] Reference sequence X55901 (SEQ ID NO: 269) represents the L genome segment of the Hantaan virus. Reference sequence NC\_004366 (SEQ ID NO: 270) represents the S genome segment of the Andes virus. Reference sequence U88410 (SEQ ID NO: 271) represents the sequence coding for the nucleoprotein gene of Crimean-Congo hemorrhagic fever virus. Reference sequence NC\_002043 (SEQ ID NO: 272) represents the L genome segment of the Rift Valley fever virus.

**Table 7: Primer Pairs for Identification of Arenaviruses**

| Primer<br>Pair<br>Number | For. Primer<br>Name                    | For. Primer<br>Sequence                                                          | For.<br>SEQ ID<br>NO: | Rev. Primer<br>Name               | Rev. Primer<br>Sequence                                                                        | Rev.<br>SEQ ID<br>NO: |
|--------------------------|----------------------------------------|----------------------------------------------------------------------------------|-----------------------|-----------------------------------|------------------------------------------------------------------------------------------------|-----------------------|
| 576                      | ARENAL_NC0042<br>97_3866_3884_<br>2P F | TTCTTGACAT <sup>a</sup> GGGT<br>CAGGG                                            | 244                   | ARENAL_NC004297<br>3979 4000 2P R | TCTGGTCAT <sup>a</sup> C <sup>a</sup> ACTA<br>GAGGTATA                                         | 257                   |
| 575                      | ARENAL_NC0042<br>97_3866_3884P<br>F    | TTCTTGATAT <sup>a</sup> GGGC<br>CAGGG                                            | 245                   | ARENAL_NC004297<br>3979 4000P R   | TCTGGTCAT <sup>a</sup> C <sup>a</sup> ACTA<br>GAAGTGTA                                         | 258                   |
| 578                      | ARENAL_NC0042<br>97_3869_3884_<br>2P F | T <sup>a</sup> TGACAT <sup>a</sup> GGGTC <sup>a</sup> A<br>GGG                   | 246                   | ARENAL_NC004297<br>3988 4005P R   | TGAGATCTGGTCAT <sup>a</sup><br>C <sup>a</sup> ACT                                              | 259                   |
| 577                      | ARENAL_NC0042<br>97_3869_3884P<br>F    | T <sup>a</sup> TGATAT <sup>a</sup> GGGCC <sup>a</sup> A<br>GGG                   | 247                   | ARENAL_NC004297<br>3988 4005P R   | TGAGATCTGGTCAT <sup>a</sup><br>C <sup>a</sup> ACT                                              | 259                   |
| 574                      | ARENAL_NC0042<br>97_3976_3995P<br>F    | TCTT <sup>a</sup> ACACCT <sup>a</sup> C <sup>a</sup> AA<br>GTGAT <sup>a</sup> GA | 248                   | ARENAL_NC004297<br>4099 4118P R   | TTAGGGC <sup>a</sup> T <sup>a</sup> GACAAA<br>CT <sup>a</sup> T <sup>a</sup> GTT               | 260                   |
| 573                      | ARENAL_NC0042<br>97_3979_4001P<br>F    | TACACTT <sup>a</sup> C <sup>a</sup> TAGTGA<br>T <sup>a</sup> GATCAGAT            | 249                   | ARENAL_NC004297<br>4099 4118P R   | TTAGGGC <sup>a</sup> T <sup>a</sup> GACAAA<br>CT <sup>a</sup> T <sup>a</sup> GTT               | 260                   |
| 570                      | ARENAL_NC0042<br>97_4105_4124P<br>F    | T <sup>a</sup> T <sup>a</sup> TGTCAGCC <sup>a</sup> C <sup>a</sup> T<br>AAAAGTGT | 250                   | ARENAL_NC004297<br>4216 4235P R   | TCT <sup>a</sup> T <sup>a</sup> T <sup>a</sup> GCACTTTA<br>CAT <sup>a</sup> T <sup>a</sup> GTG | 261                   |
| 571                      | ARENAL_NC0042<br>97_4105_4124P<br>F    | T <sup>a</sup> T <sup>a</sup> TGTCAGCC <sup>a</sup> C <sup>a</sup> T<br>AAAAGTGT | 251                   | ARENAL_NC004297<br>4201 4219 R    | TGTGTAGCGCTGCAG<br>CAAC                                                                        | 262                   |
| 572                      | ARENAL_NC0042<br>97_4812_4829P<br>F    | TAACAAAT <sup>a</sup> C <sup>a</sup> AGCAT<br><sup>a</sup> T <sup>a</sup> CCA    | 252                   | ARENAL_NC004297<br>4841 4857P R   | TC <sup>a</sup> C <sup>a</sup> T <sup>a</sup> ATAAAGC <sup>a</sup> C <sup>a</sup><br>AGATG     | 263                   |
| 582                      | ARENAN_NC0042<br>96_474_494_2_<br>F    | TGGTGTGTGAGAGT<br>CTGGGA                                                         | 253                   | ARENAN_NC002496<br>520 540 2 R    | TGGCATTGACCCAAA<br>CTGGTT                                                                      | 264                   |
| 581                      | ARENAN_NC0042<br>96_474_494 F          | TGGTGTGTGAAGGT<br>CTGGGA                                                         | 254                   | ARENAN_NC002496<br>520 540 R      | TGGCATTGACCCGAA<br>CTGATT                                                                      | 265                   |
| 580                      | ARENAN_NC0042<br>96_937_953 2P         | T <sup>a</sup> C <sup>a</sup> AGGTGAAGGTT <sup>a</sup><br>GGC <sup>a</sup> C     | 255                   | ARENAN_NC002496<br>982 1002 2 R   | TGTGTGTGCCCAAGC<br>CCTTCC                                                                      | 266                   |

- 58 -

|     |                                |                                                                              |     |                               |                             |     |
|-----|--------------------------------|------------------------------------------------------------------------------|-----|-------------------------------|-----------------------------|-----|
|     | F                              |                                                                              |     |                               |                             |     |
| 579 | ARENAN_NC0042<br>96 937 953P F | T <sup>a</sup> C <sup>a</sup> AGGTGATGGAT <sup>a</sup><br>GGC <sup>a</sup> C | 256 | ARENAN_NC002496<br>982 1002 R | TGTG TTGTCCCAAGC<br>TCTC CC | 267 |

[0195] Reference sequence NC\_004297 (SEQ ID NO: 273) represents the L genome segment of the Lassa virus. Reference sequence NC\_004296 (SEQ ID NO: 274) represents the S genome segment of the Lassa virus.

### Example 2: One-Step RT-PCR of RNA Virus Samples

[0196] RNA was isolated from virus-containing samples according to methods well known in the art. To generate bioagent identifying amplicons for RNA viruses, a one-step RT-PCR protocol was developed. All RT-PCR reactions were assembled in 50 µl reactions in the 96 well microtiter plate format using a Packard MPII liquid handling robotic platform and MJ Dyad<sup>®</sup> thermocyclers (MJ research, Waltham, MA). The RT-PCR reaction consisted of 4 units of Amplitaq Gold<sup>®</sup>, 1.5x buffer II (Applied Biosystems, Foster City, CA), 1.5 mM MgCl<sub>2</sub>, 0.4 M betaine, 10 mM DTT, 20 mM sorbitol, 50 ng random primers (Invitrogen, Carlsbad, CA), 1.2 units Supersin (Ambion, Austin, TX), 100 ng polyA DNA, 2 units Superscript<sup>™</sup> III (Invitrogen, Carlsbad, CA), 400 ng T4 Gene 32 Protein (Roche Applied Science, Indianapolis, IN), 800 µM dNTP mix, and 250 nM of each primer.

[0197] The following RT-PCR conditions were used to amplify the sequences used for mass spectrometry analysis: 60°C for 5 minutes, 4°C for 10 minutes, 55°C for 45 minutes, 95°C for 10 minutes followed by 8 cycles of 95°C for 30 seconds, 48°C for 30 seconds, and 72°C for 30 seconds, with the 48°C annealing temperature increased 0.9°C after each cycle. The PCR reaction was then continued for 37 additional cycles of 95°C for 15 seconds, 56°C for 20 seconds, and 72°C for 20 seconds. The reaction concluded with 2 minutes at 72°C.

### Example 3: Solution Capture Purification of PCR Products for Mass Spectrometry with Ion Exchange Resin-Magnetic Beads

[0198] For solution capture of nucleic acids with ion exchange resin linked to magnetic beads, 25 µl of a 2.5 mg/mL suspension of BioClon amine terminated supraparamagnetic beads were added to 25 to 50 µl of a PCR (or RT-PCR) reaction containing approximately 10 pM of a typical PCR amplification product. The above suspension was mixed for approximately 5 minutes by vortexing or pipetting, after which the liquid was removed after using a magnetic separator. The beads containing bound PCR amplification product were then washed 3x with

- 59 -

50mM ammonium bicarbonate/50% MeOH or 100mM ammonium bicarbonate/50% MeOH, followed by three more washes with 50% MeOH. The bound PCR amplicon was eluted with 25mM piperidine, 25mM imidazole, 35% MeOH, plus peptide calibration standards.

**Example 4: Mass Spectrometry and Base Composition Analysis**

[0199] The ESI-FTICR mass spectrometer is based on a Bruker Daltonics (Billerica, MA) Apex II 70e electrospray ionization Fourier transform ion cyclotron resonance mass spectrometer that employs an actively shielded 7 Tesla superconducting magnet. The active shielding constrains the majority of the fringing magnetic field from the superconducting magnet to a relatively small volume. Thus, components that might be adversely affected by stray magnetic fields, such as CRT monitors, robotic components, and other electronics, can operate in close proximity to the FTICR spectrometer. All aspects of pulse sequence control and data acquisition were performed on a 600 MHz Pentium II data station running Bruker's Xmass<sup>TM</sup> software under Windows NT 4.0 operating system. Sample aliquots, typically 15  $\mu$ l, were extracted directly from 96-well microtiter plates using a CTC HTS PAL autosampler (LEAP Technologies, Carrboro, NC) triggered by the FTICR data station. Samples were injected directly into a 10  $\mu$ l sample loop integrated with a fluidics handling system that supplies the 100  $\mu$ l/hr flow rate to the ESI source. Ions were formed via electrospray ionization in a modified Analytica<sup>TM</sup> (Branford, CT) source employing an off axis, grounded electrospray probe positioned approximately 1.5 cm from the metalized terminus of a glass desolvation capillary. The atmospheric pressure end of the glass capillary was biased at 6000 V relative to the ESI needle during data acquisition. A counter-current flow of dry N<sub>2</sub> was employed to assist in the desolvation process. Ions were accumulated in an external ion reservoir comprised of an rf-only hexapole, a skimmer cone, and an auxiliary gate electrode, prior to injection into the trapped ion cell where they were mass analyzed. Ionization duty cycles > 99% were achieved by simultaneously accumulating ions in the external ion reservoir during ion detection. Each detection event consisted of 1M data points digitized over 2.3 s. To improve the signal-to-noise ratio (S/N), 32 scans were co-added for a total data acquisition time of 74 s.

[0200] The ESI-TOF mass spectrometer is based on a Bruker Daltonics MicroTOF<sup>TM</sup>. Ions from the ESI source undergo orthogonal ion extraction and are focused in a reflectron prior to detection. The TOF and FTICR are equipped with the same automated sample handling and fluidics described above. Ions are formed in the standard MicroTOF<sup>TM</sup> ESI source that is equipped with the same off-axis sprayer and glass capillary as the FTICR ESI source.

- 60 -

Consequently, source conditions were the same as those described above. External ion accumulation was also employed to improve ionization duty cycle during data acquisition. Each detection event on the TOF was comprised of 75,000 data points digitized over 75  $\mu$ s.

[0201] The sample delivery scheme allows sample aliquots to be rapidly injected into the electrospray source at high flow rate and subsequently be electrosprayed at a much lower flow rate for improved ESI sensitivity. Prior to injecting a sample, a bolus of buffer was injected at a high flow rate to rinse the transfer line and spray needle to avoid sample contamination/carryover. Following the rinse step, the autosampler injected the next sample and the flow rate was switched to low flow. Following a brief equilibration delay, data acquisition commenced. As spectra were co-added, the autosampler continued rinsing the syringe and picking up buffer to rinse the injector and sample transfer line. In general, two syringe rinses and one injector rinse were required to minimize sample carryover. During a routine screening protocol a new sample mixture was injected every 106 seconds. More recently a fast wash station for the syringe needle has been implemented which, when combined with shorter acquisition times, facilitates the acquisition of mass spectra at a rate of just under one spectrum/minute.

[0202] Raw mass spectra were post-calibrated with an internal mass standard and deconvoluted to monoisotopic molecular masses. Unambiguous base compositions were derived from the exact mass measurements of the complementary single-stranded oligonucleotides. Quantitative results are obtained by comparing the peak heights with an internal PCR calibration standard present in every PCR well at 500 molecules per well. Calibration methods are commonly owned and disclosed in U.S. Provisional Patent Application Serial No. 60/545,425.

**Example 5: *De Novo* Determination of Base Composition of Amplification Products using Molecular Mass Modified Deoxynucleotide Triphosphates**

[0203] Because the molecular masses of the four natural nucleobases have a relatively narrow molecular mass range (A = 313.058, G = 329.052, C = 289.046, T = 304.046 – See Table 8), a persistent source of ambiguity in assignment of base composition can occur as follows: two nucleic acid strands having different base composition may have a difference of about 1 Da when the base composition difference between the two strands is G  $\leftrightarrow$  A (-15.994) combined with C  $\leftrightarrow$  T (+15.000). For example, one 99-mer nucleic acid strand having a base composition

of  $A_{27}G_{30}C_{21}T_{21}$  has a theoretical molecular mass of 30779.058 while another 99-mer nucleic acid strand having a base composition of  $A_{26}G_{31}C_{22}T_{20}$  has a theoretical molecular mass of 30780.052. A 1 Da difference in molecular mass may be within the experimental error of a molecular mass measurement and thus, the relatively narrow molecular mass range of the four natural nucleobases imposes an uncertainty factor.

[0204] The present invention provides for a means for removing this theoretical 1 Da uncertainty factor through amplification of a nucleic acid with one mass-tagged nucleobase and three natural nucleobases. The term “nucleobase” as used herein is synonymous with other terms in use in the art including “nucleotide,” “deoxynucleotide,” “nucleotide residue,” “deoxynucleotide residue,” “nucleotide triphosphate (NTP),” or deoxynucleotide triphosphate (dNTP).

[0205] Addition of significant mass to one of the 4 nucleobases (dNTPs) in an amplification reaction, or in the primers themselves, will result in a significant difference in mass of the resulting amplification product (significantly greater than 1 Da) arising from ambiguities arising from the  $G \leftrightarrow A$  combined with  $C \leftrightarrow T$  event (Table 8). Thus, the same the  $G \leftrightarrow A$  (-15.994) event combined with 5-Iodo- $C \leftrightarrow T$  (-110.900) event would result in a molecular mass difference of 126.894. If the molecular mass of the base composition  $A_{27}G_{30}$  **5-Iodo- $C_{21}T_{21}$**  (33422.958) is compared with  $A_{26}G_{31}$  **5-Iodo- $C_{22}T_{20}$** , (33549.852) the theoretical molecular mass difference is +126.894. The experimental error of a molecular mass measurement is not significant with regard to this molecular mass difference. Furthermore, the only base composition consistent with a measured molecular mass of the 99-mer nucleic acid is  $A_{27}G_{30}$  **5-Iodo- $C_{21}T_{21}$** . In contrast, the analogous amplification without the mass tag has 18 possible base compositions.

**Table 8: Molecular Masses of Natural Nucleobases and the Mass-Modified Nucleobase 5-Iodo-C and Molecular Mass Differences Resulting from Transitions**

| Nucleobase | Molecular Mass | Transition   | $\Delta$ Molecular Mass |
|------------|----------------|--------------|-------------------------|
| A          | 313.058        | A-->T        | -9.012                  |
| A          | 313.058        | A-->C        | -24.012                 |
| A          | 313.058        | A-->5-Iodo-C | 101.888                 |
| A          | 313.058        | A-->G        | 15.994                  |
| T          | 304.046        | T-->A        | 9.012                   |
| T          | 304.046        | T-->C        | -15.000                 |
| T          | 304.046        | T-->5-Iodo-C | 110.900                 |
| T          | 304.046        | T-->G        | 25.006                  |
| C          | 289.046        | C-->A        | 24.012                  |

- 62 -

|          |         |              |          |
|----------|---------|--------------|----------|
| C        | 289.046 | C-->T        | 15.000   |
| C        | 289.046 | C-->G        | 40.006   |
| 5-Iodo-C | 414.946 | 5-Iodo-C-->A | -101.888 |
| 5-Iodo-C | 414.946 | 5-Iodo-C-->T | -110.900 |
| 5-Iodo-C | 414.946 | 5-Iodo-C-->G | -85.894  |
| G        | 329.052 | G-->A        | -15.994  |
| G        | 329.052 | G-->T        | -25.006  |
| G        | 329.052 | G-->C        | -40.006  |
| G        | 329.052 | G-->5-Iodo-C | 85.894   |

**Example 6: Data Processing**

[0206] Mass spectra of bioagent-identifying amplicons are analyzed independently using *e.g.*, a maximum-likelihood processor, such as is widely used in radar signal processing. This processor, referred to as GenX<sup>TM</sup>, first makes maximum likelihood estimates of the input to the mass spectrometer for each primer by running matched filters for each base composition aggregate on the input data. This includes the GenX<sup>TM</sup> response to a calibrant for each primer.

[0207] The algorithm emphasizes performance predictions culminating in probability-of-detection versus probability-of-false-alarm plots for conditions involving complex backgrounds of naturally occurring organisms and environmental contaminants. Matched filters consist of *a priori* expectations of signal values given the set of primers used for each of the bioagents. A genomic sequence database is used to define the mass base count matched filters. The database contains the sequences of known bacterial bioagents and includes threat organisms as well as benign background organisms. The latter is used to estimate and subtract the spectral signature produced by the background organisms. A maximum likelihood detection of known background organisms is implemented using matched filters and a running-sum estimate of the noise covariance. Background signal strengths are estimated and used along with the matched filters to form signatures which are then subtracted. the maximum likelihood process is applied to this "cleaned up" data in a similar manner employing matched filters for the organisms and a running-sum estimate of the noise-covariance for the cleaned up data.

[0208] The amplitudes of all base compositions of bioagent-identifying amplicons for each primer are calibrated and a final maximum likelihood amplitude estimate per organism is made based upon the multiple single primer estimates. Models of all system noise are factored into this two-stage maximum likelihood calculation. The processor reports the number of molecules of each base composition contained in the spectra. The quantity of amplification product

- 63 -

corresponding to the appropriate primer set is reported as well as the quantities of primers remaining upon completion of the amplification reaction.

[0209] Base count blurring can be carried out as follows. "Electronic PCR" can be conducted on nucleotide sequences of the desired bioagents to obtain the different expected base counts that could be obtained for each primer pair.

In one illustrative

embodiment, one or more spreadsheets, such as Microsoft Excel workbooks contains a plurality of worksheets. First in this example, there is a worksheet with a name similar to the workbook name; this worksheet contains the raw electronic PCR data. Second, there is a worksheet named "filtered bioagents base count" that contains bioagent name and base count; there is a separate record for each strain after removing sequences that are not identified with a genus and species and removing all sequences for bioagents with less than 10 strains. Third, there is a worksheet, "Sheet1" that contains the frequency of substitutions, insertions, or deletions for this primer pair. This data is generated by first creating a pivot table from the data in the "filtered bioagents base count" worksheet and then executing an Excel VBA macro. The macro creates a table of differences in base counts for bioagents of the same species, but different strains. One of ordinary skill in the art may understand additional pathways for obtaining similar table differences without undo experimentation.

[0210] Application of an exemplary script, involves the user defining a threshold that specifies the fraction of the strains that are represented by the reference set of base counts for each bioagent. The reference set of base counts for each bioagent may contain as many different base counts as are needed to meet or exceed the threshold. The set of reference base counts is defined by taking the most abundant strain's base type composition and adding it to the reference set and then the next most abundant strain's base type composition is added until the threshold is met or exceeded. The current set of data were obtained using a threshold of 55%, which was obtained empirically.

[0211] For each base count not included in the reference base count set for that bioagent, the script then proceeds to determine the manner in which the current base count differs from each of the base counts in the reference set. This difference may be represented as a combination of substitutions,  $S_i=X_i$ , and insertions,  $I_i=Y_i$ , or deletions,  $D_i=Z_i$ . If there is more than one reference base count, then the reported difference is chosen using rules that aim to minimize the

- 64 -

number of changes and, in instances with the same number of changes, minimize the number of insertions or deletions. Therefore, the primary rule is to identify the difference with the minimum sum ( $X_i + Y_i$ ) or ( $X_i + Z_i$ ), *e.g.*, one insertion rather than two substitutions. If there are two or more differences with the minimum sum, then the one that will be reported is the one that contains the most substitutions.

**[0212]** Differences between a base count and a reference composition are categorized as either one, two, or more substitutions, one, two, or more insertions, one, two, or more deletions, and combinations of substitutions and insertions or deletions. The different types of changes and their probabilities of occurrence have been delineated in U.S. Patent Application Publication No. 2004209260 (U.S. Application Serial No. 10/418,514).

#### **Example 7: Identification of Five Different Strains of Filoviruses**

**[0213]** Four primer pairs from Table 5 – primer pair nos. 853 (SEQ ID NOs: 129:164), 856 (SEQ ID NOs: 134:169), 858 (SEQ ID NOs: 124:159) and 864 (SEQ ID NOs: 138:174) were selected as candidate primer pairs for providing broad coverage of all known viral bioagents in the filoviridae family after amplification tests of 24 primer pairs wherein efficiency of primer pair amplification was assessed by gel electrophoresis. Each of these four primer pairs targets the L polymerase gene region. Samples of isolates of Zaire Ebola virus (Mayinga strain), Sudan Ebola virus (Boniface strain), Reston Ebola virus (Reston strain), and two isolates of Marburg virus (M/Kenya/Kitum/Cave/1987/Ravn strain and Voegel strain) were obtained from the Center for Disease Control (CDC). RNA was isolated and reverse transcribed from these isolate samples according to Example 2. For each different reaction wherein a different primer pair used (primer pair numbers 853, 856, 858 and 864), the resulting cDNA was diluted by a factor of  $10^{-3}$  to  $10^{-6}$  and 100 copies of a calibration polynucleotide (SEQ ID NO: 275) contained within the pCR Blunt® vector (Invitrogen, Carlsbad, CA) was spiked into the sample. The calibration polynucleotide is based upon a portion of sequence of the Zaire Ebola virus (Mayinga) genome (SEQ ID NO: 268) and contains a series of deletions 5 nucleobases in length which, for each amplification product produced by primer pair numbers 853, 856, 858 and 864, provide enough of a difference in molecular mass to distinguish each calibration amplicon from the corresponding filovirus identifying amplicon. The 5 nucleobase deletions are located at the following coordinates with respect to the reference sequence (SEQ ID NO: 268): 15339-15343, 15441-15445, 15583-15587, 15641-15645, and 15772-15776.

[0214] The viral isolate cDNA and the calibrant were amplified and amplification products were purified by magnetic solution capture according to Example 3 followed by mass spectrometric analysis according to Example 4. Base compositions were deconvolved from the molecular masses of the filovirus identifying amplicons and are shown in Table 9 along with the expected base compositions based on known sequence information. It should be noted that primer pair number 858 was not expected to prime the Sudan Ebola virus and, as expected, an amplification product was not observed.

[0215] This example indicates that the four primer pairs investigated are functional in their intended purpose for producing filovirus identifying amplicons with base compositions that can identify different filovirus strains.

**Table 9: Expected and Observed Base Compositions of Filovirus Identifying Amplicons Produced with Primer Pair Nos: 853 (SEQ ID NOs: 129:164), 856 (SEQ ID NOs: 134:169), 858 (SEQ ID NOs: 124:159) and 864 (SEQ ID NOs: 138:174)**

| Virus              | Strain                       | Sequence Available | Primer Pair | Expected Base Composition<br>[A G C T] | Observed Base Composition<br>[A G C T] |
|--------------------|------------------------------|--------------------|-------------|----------------------------------------|----------------------------------------|
| Zaire Ebola virus  | Mayinga                      | Yes                | 853         | [20 19 14 26]                          | [20 19 14 26]                          |
| Sudan Ebola virus  | Boniface                     | Yes                | 853         | [18 19 15 27]                          | [18 19 15 27]                          |
| Reston Ebola virus | Reston                       | Yes                | 853         | [18 20 13 28]                          | [18 20 13 28]                          |
| Marburg virus      | M/Kenya/Kitum/Cave/1987/Ravn | No                 | 853         | -                                      | [19 18 13 29]                          |
| Marburg virus      | Voege                        | No                 | 853         | -                                      | [19 18 13 29]                          |
| Calibrant          | Based on Ebola Virus Zaire   | Yes                | 853         | [17 19 13 25]                          | [17 19 13 25]                          |
| Zaire Ebola virus  | Mayinga                      | Yes                | 856         | [50 35 40 32]                          | [50 35 40 32]                          |
| Sudan Ebola virus  | Boniface                     | Yes                | 856         | [47 36 34 40]                          | [47 36 34 40]                          |
| Reston Ebola virus | Reston                       | Yes                | 856         | [48 35 36 38]                          | [48 35 36 38]                          |
| Marburg virus      | M/Kenya/Kitum/Cave/1987/Ravn | No                 | 856         | -                                      | [41 35 30 51]                          |
| Marburg virus      | Voege                        | No                 | 856         | -                                      | [50 33 34 40]                          |
| Calibrant          | Based on Ebola Virus Zaire   | Yes                | 856         | [49 34 38 31]                          | [49 34 38 31]                          |
| Zaire Ebola virus  | Mayinga                      | Yes                | 858         | [32 24 22 25]                          | [32 24 22 25]                          |
| Sudan Ebola virus  | Boniface                     | Yes                | 858         | Amplification not expected             | Amplification not observed             |
| Reston Ebola       | Reston                       | Yes                | 858         | [30 25 22 26]                          | [30 25 22 26]                          |

- 66 -

| virus              |                                  |     |     |               |               |
|--------------------|----------------------------------|-----|-----|---------------|---------------|
| Marburg virus      | M/Kenya/Kitum/Cave/<br>1987/Ravn | No  | 858 | -             | [34 24 20 25] |
| Marburg virus      | Voegel                           | No  | 858 | -             | [34 24 22 23] |
| Calibrant          | Based on Ebola Virus<br>Zaire    | Yes | 858 | [30 23 21 24] | [30 23 21 24] |
| Zaire Ebola virus  | Mayinga                          | Yes | 864 | [29 22 14 33] | [29 22 14 33] |
| Sudan Ebola virus  | Boniface                         | Yes | 864 | [32 23 17 26] | [32 23 17 26] |
| Reston Ebola virus | Reston                           | Yes | 864 | [31 22 16 29] | [31 22 16 29] |
| Marburg virus      | M/Kenya/Kitum/Cave/<br>1987/Ravn | No  | 864 | -             | [36 20 16 26] |
| Marburg virus      | Voegel                           | No  | 864 | -             | [38 17 15 28] |
| Calibrant          | Based on Ebola Virus<br>Zaire    | Yes | 864 | [28 20 13 32] | [28 20 13 32] |

[0216] Various modifications of the invention, in addition to those described herein, will be apparent to those skilled in the art from the foregoing description. Such modifications are also intended to fall within the scope of the appended claims.

Those skilled in the art will appreciate that numerous changes and modifications may be made to the embodiments of the invention and that such changes and modifications may be made without departing from the spirit of the invention. It is therefore intended that the appended claims cover all such equivalent variations as fall within the true spirit and scope of the invention.

# **DEMANDES OU BREVETS VOLUMINEUX**

**LA PRÉSENTE PARTIE DE CETTE DEMANDE OU CE BREVETS  
COMPREND PLUS D'UN TOME.**

**CECI EST LE TOME \_\_1\_\_ DE \_\_2\_\_**

NOTE: Pour les tomes additionels, veuillez contacter le Bureau Canadien des Brevets.

---

# **JUMBO APPLICATIONS / PATENTS**

**THIS SECTION OF THE APPLICATION / PATENT CONTAINS MORE  
THAN ONE VOLUME.**

**THIS IS VOLUME \_\_1\_\_ OF \_\_2\_\_**

NOTE: For additional volumes please contact the Canadian Patent Office.

**Claims:**

1. An oligonucleotide primer for identification of a filovirus oligonucleotide comprising 100% sequence identity with SEQ ID NO: 129.
2. A composition comprising the primer of claim 1 and a second primer wherein said second primer comprises an oligonucleotide primer comprising 100% sequence identity with SEQ ID NO: 164.
3. The composition of claim 2 wherein either or both of said primers comprises at least one modified nucleobase.
4. The composition of claim 3 wherein said modified nucleobase is 5-propynyluracil or 5-propynylcytosine.
5. The composition of claim 2 wherein either or both of said primers comprises at least one universal nucleobase.
6. The composition of claim 6 wherein said universal nucleobase is inosine.
7. The composition of claim 2 wherein either or both of said primers further comprises a non-templated T residue on the 5'-end.
8. The composition of claim 2 wherein either or both of said primers comprises at least one non-template tag, wherein said non-template tag comprises at least three consecutive A or T nucleotide residues which are not complementary to a primer template.
9. The composition of claim 2 wherein either or both of said primers comprises at least one molecular mass modifying tag.

10. A kit for amplifying a filovirus nucleic acid comprising an oligonucleotide primer comprising 100%sequence identity with SEQ ID NO: 129 and an oligonucleotide primer comprising 100%sequence identity with SEQ ID NO:164.

11. The kit of claim 10 further comprising at least one ion exchange resin linked to magnetic beads.

12. A method for identification of an unknown filovirus comprising:

- amplifying nucleic acid from said filovirus using the composition of claim 2 to obtain an amplification product;

- determining the molecular mass of said amplification product by mass spectrometry;

- determining the base composition of said amplification product from said molecular mass; and

- comparing said molecular mass or base composition with a plurality of molecular masses or base compositions of known filoviral bioagent identifying amplicons amplified using the composition of claim 2, wherein a match between said molecular mass or base composition and a member of said plurality of molecular masses or base compositions identifies said unknown filovirus.

13. A method of determining the presence or absence of a filovirus in a sample comprising:

- amplifying nucleic acid from said sample using the composition of claim 2 to obtain an amplification product;

- determining the molecular mass of said amplification product by mass spectrometry;

- determining the base composition of said amplification product from said molecular mass; and

- comparing said molecular mass or base composition of said amplification product with known molecular masses or base compositions of one or more known filoviral bioagent identifying amplicons amplified using the composition of claim 2, wherein a match between said molecular mass or base composition of said amplification product and the molecular mass or base composition of one or more known filoviral bioagent identifying amplicons indicates the presence of said filovirus in said sample, and wherein the absence of a match between said

molecular mass or base composition of said amplification product and the molecular mass or base composition of one or more known filoviral bioagent identifying amplicons indicates the absence of said filovirus in said sample.

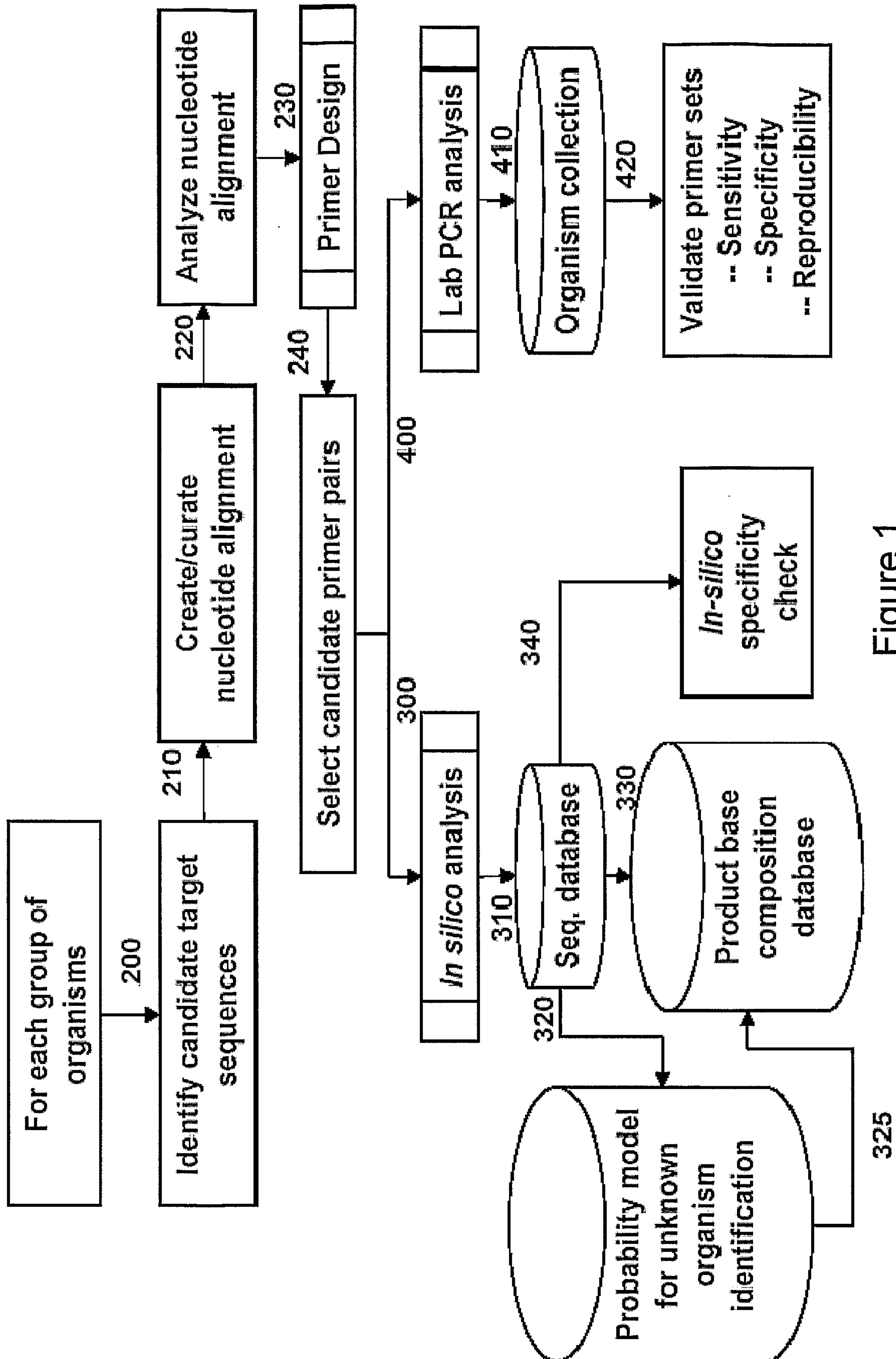


Figure 1

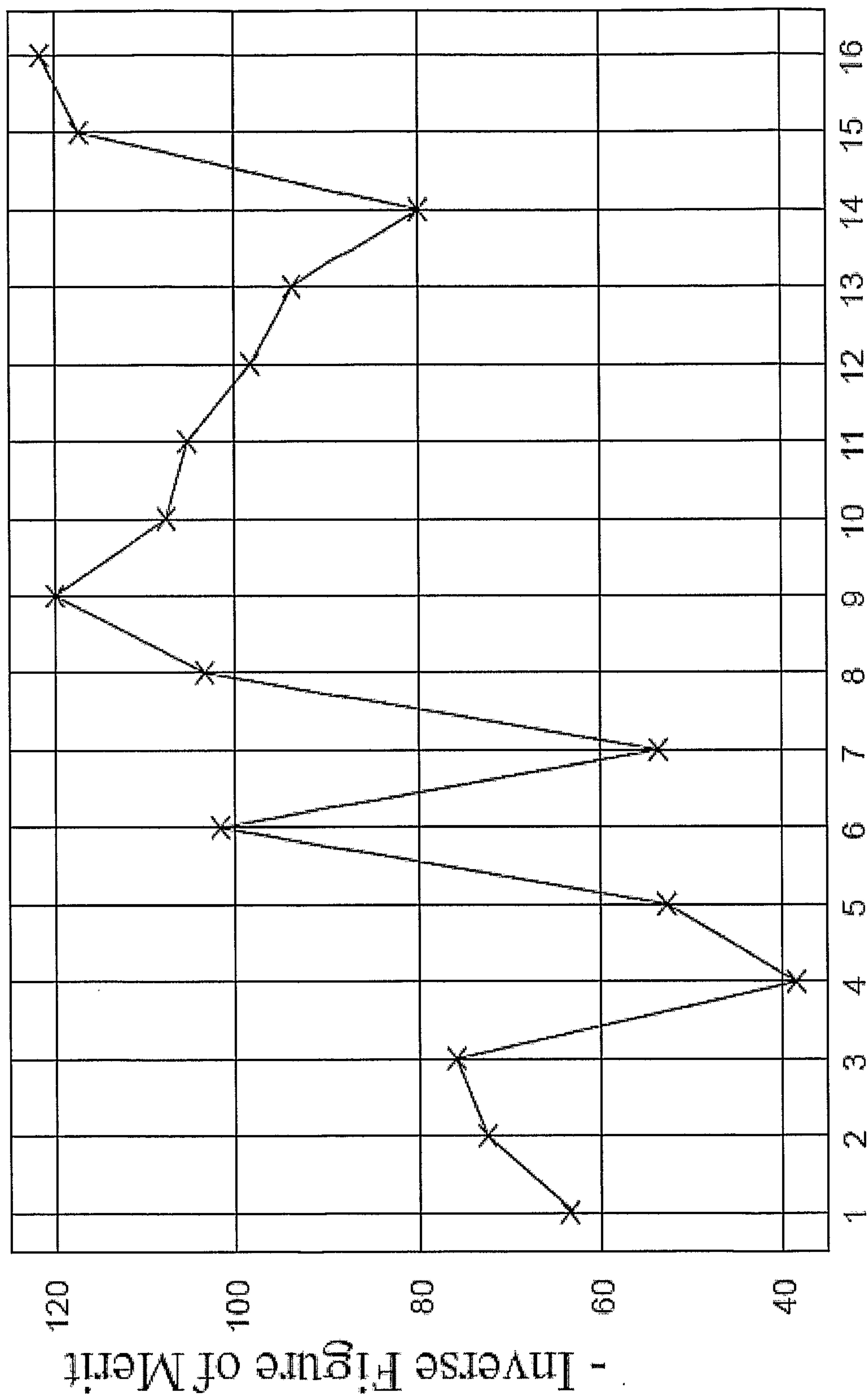


Figure 2 Primer Number

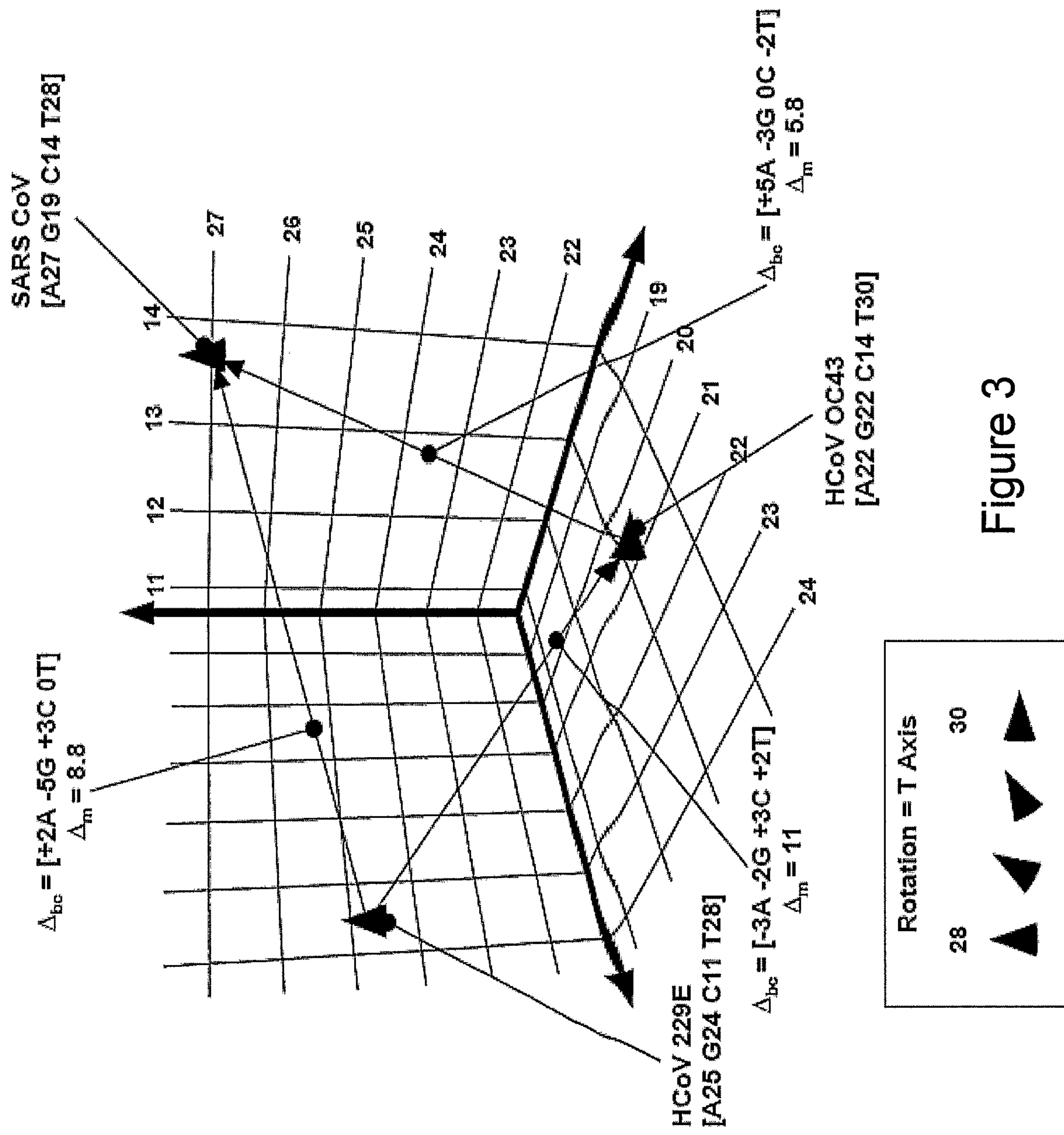


Figure 3

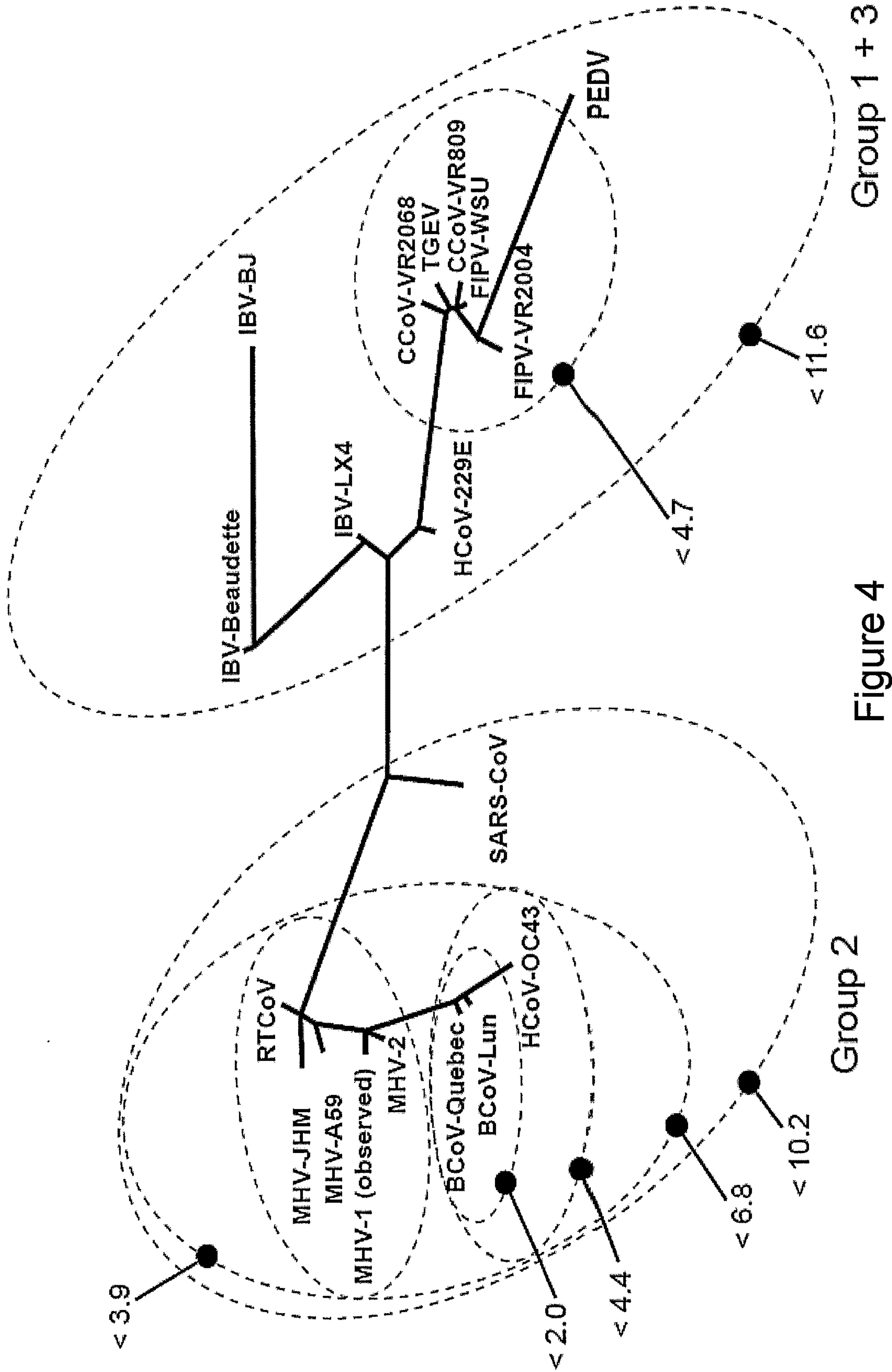


Figure 4

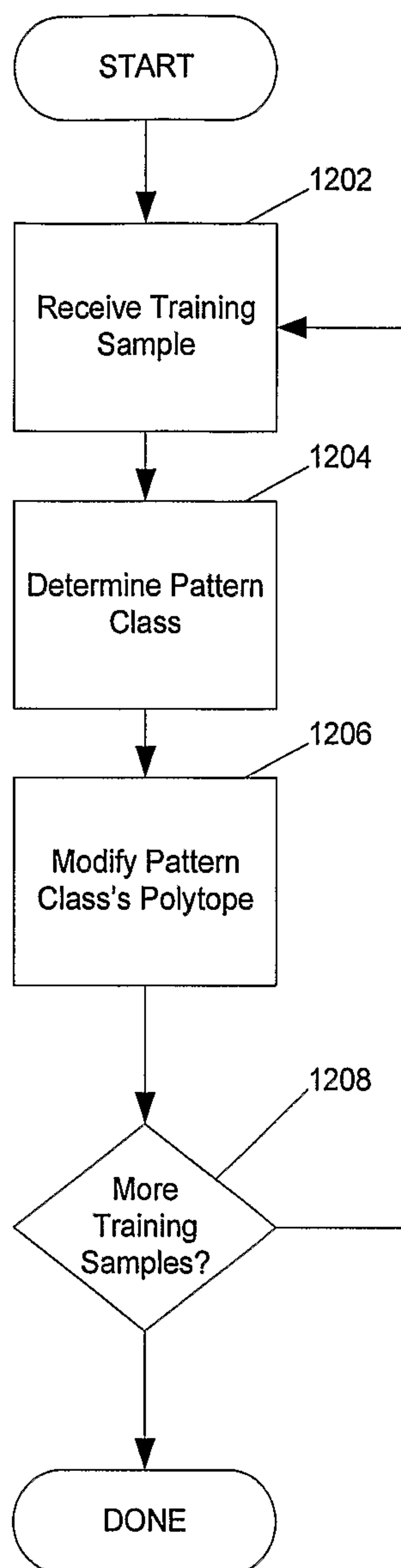


Figure 5A

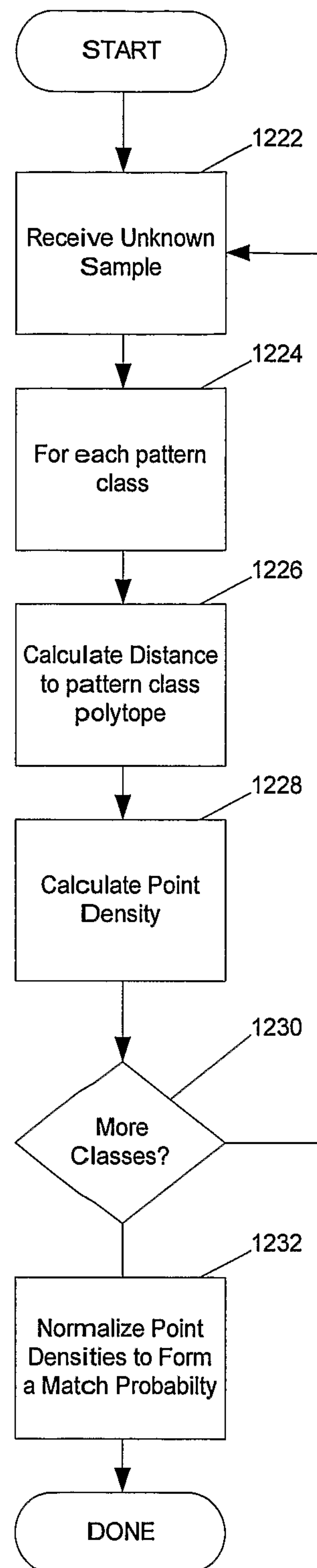


Figure 5B

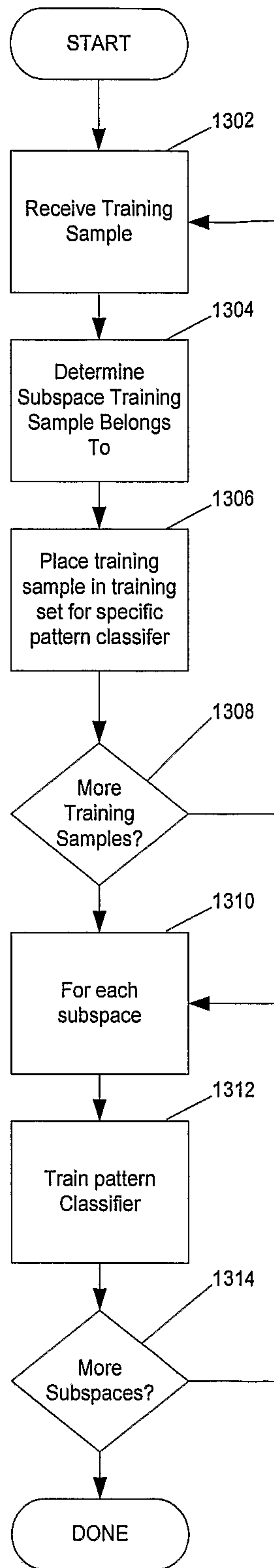


Figure 6A

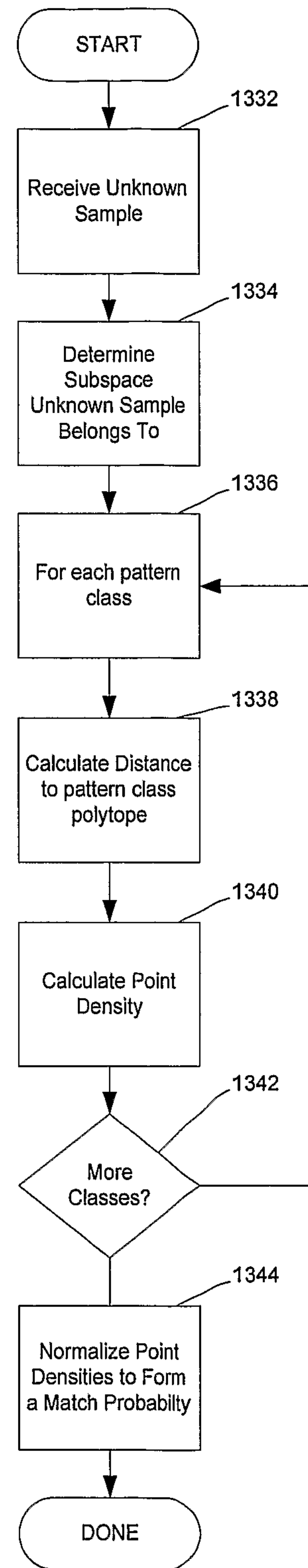


Figure 6B

7/18

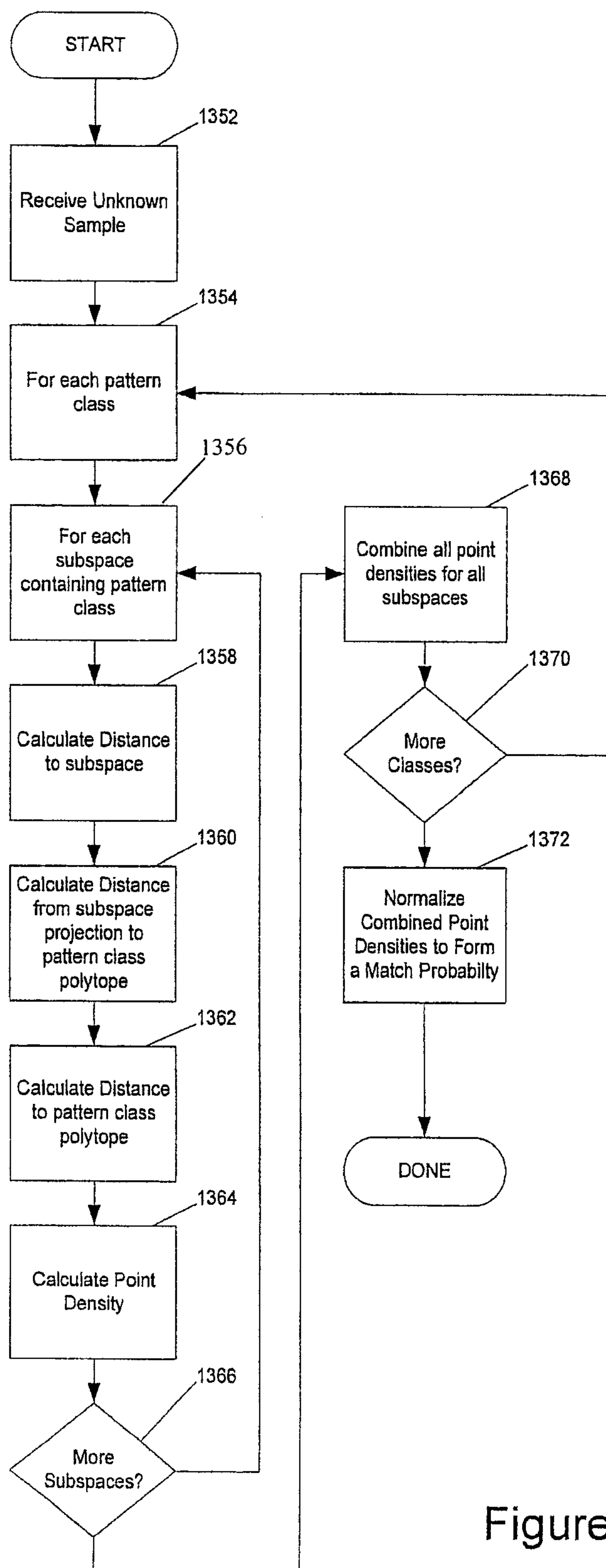
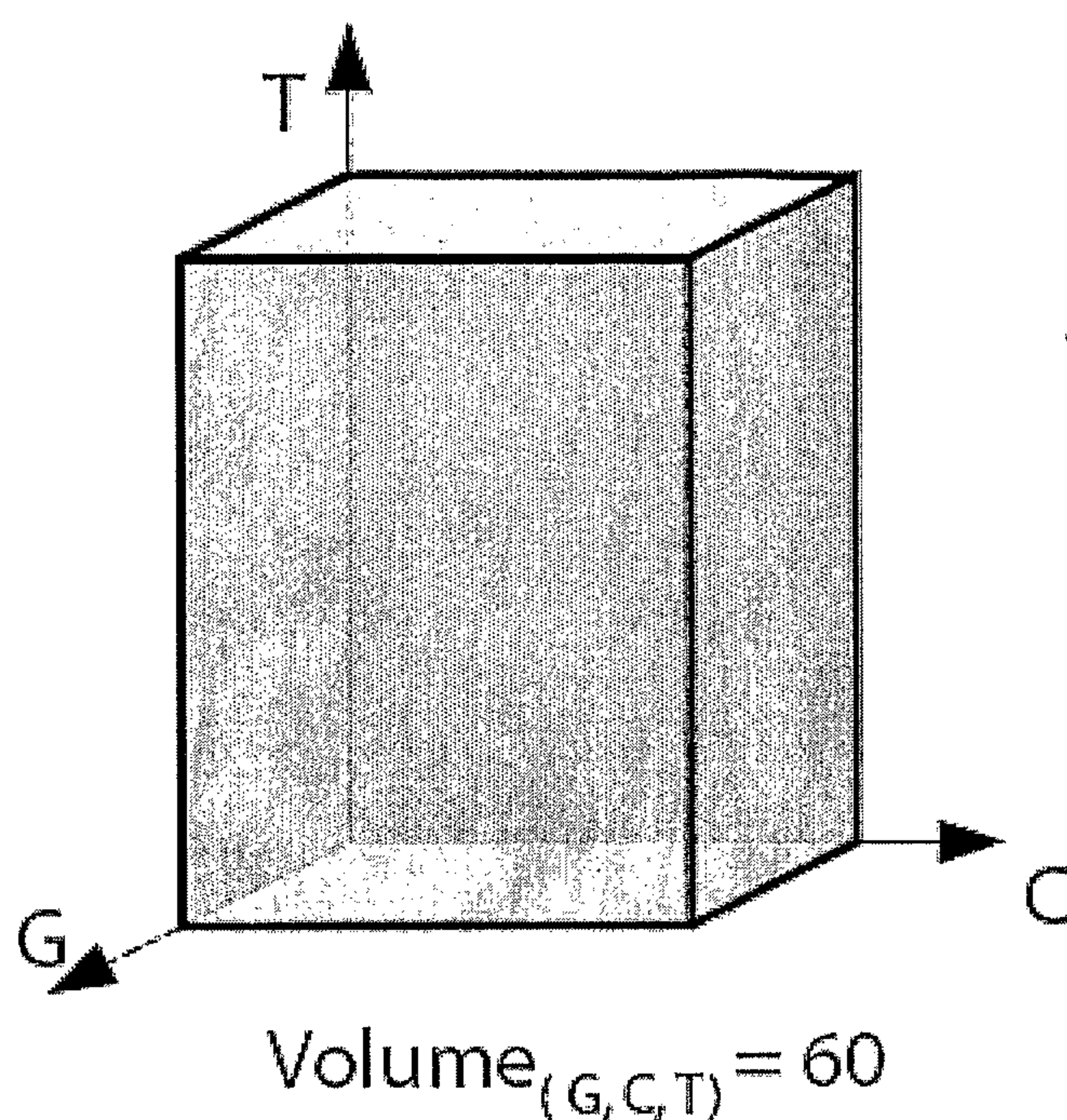


Figure 6C



$$A+G+C+T = 56$$

$$15 \leq A \leq 17$$

$$16 \leq G \leq 18$$

$$13 \leq C \leq 16$$

$$7 \leq T \leq 11$$

|                                       | A  | G  | C  | T  |
|---------------------------------------|----|----|----|----|
| <i>Neisseria gonorrhoeae</i> B 5025   | 16 | 16 | 13 | 11 |
| <i>Neisseria weaveri</i>              | 16 | 16 | 13 | 11 |
| <i>Formivibrio citricus</i>           | 17 | 16 | 16 | 7  |
| <i>Aquaspirillum delicatum</i>        | 15 | 17 | 15 | 9  |
| <i>Aquaspirillum sinuosum</i>         | 15 | 17 | 15 | 9  |
| <i>Aquaspirillum gracile</i>          | 15 | 17 | 16 | 8  |
| <i>Microvirgula aerodenitrificans</i> | 16 | 18 | 14 | 8  |

Figure 7A

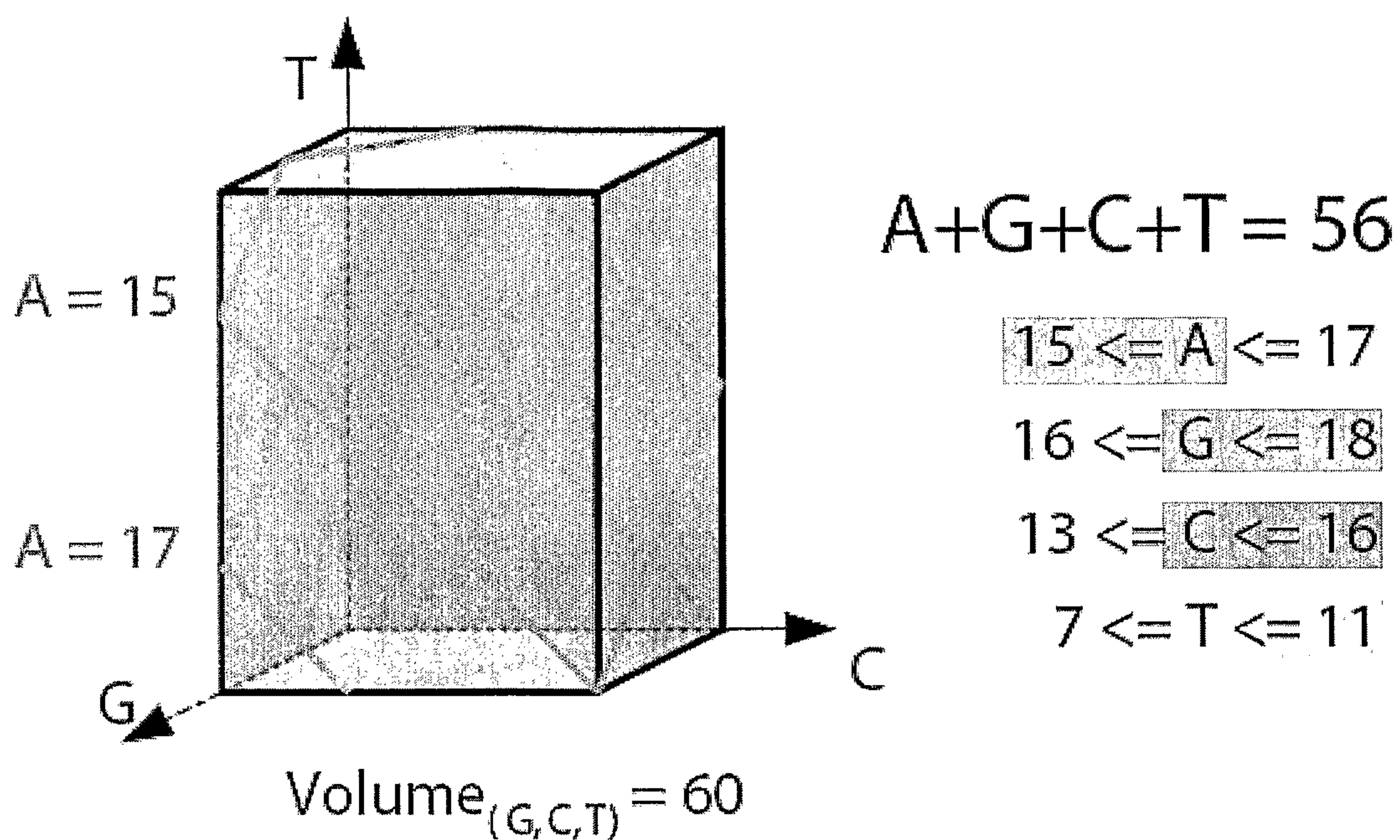


Figure 7B

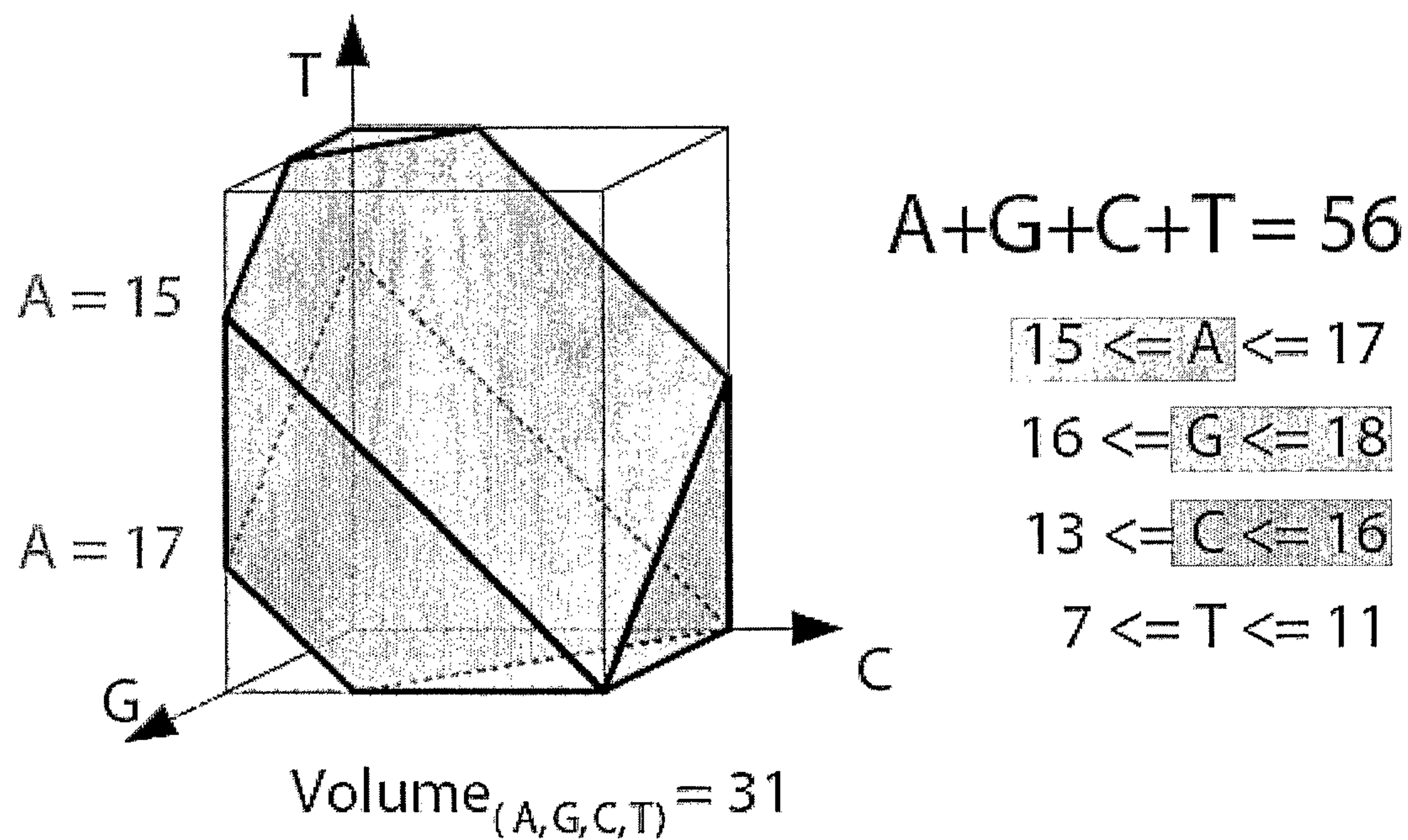


Figure 7C

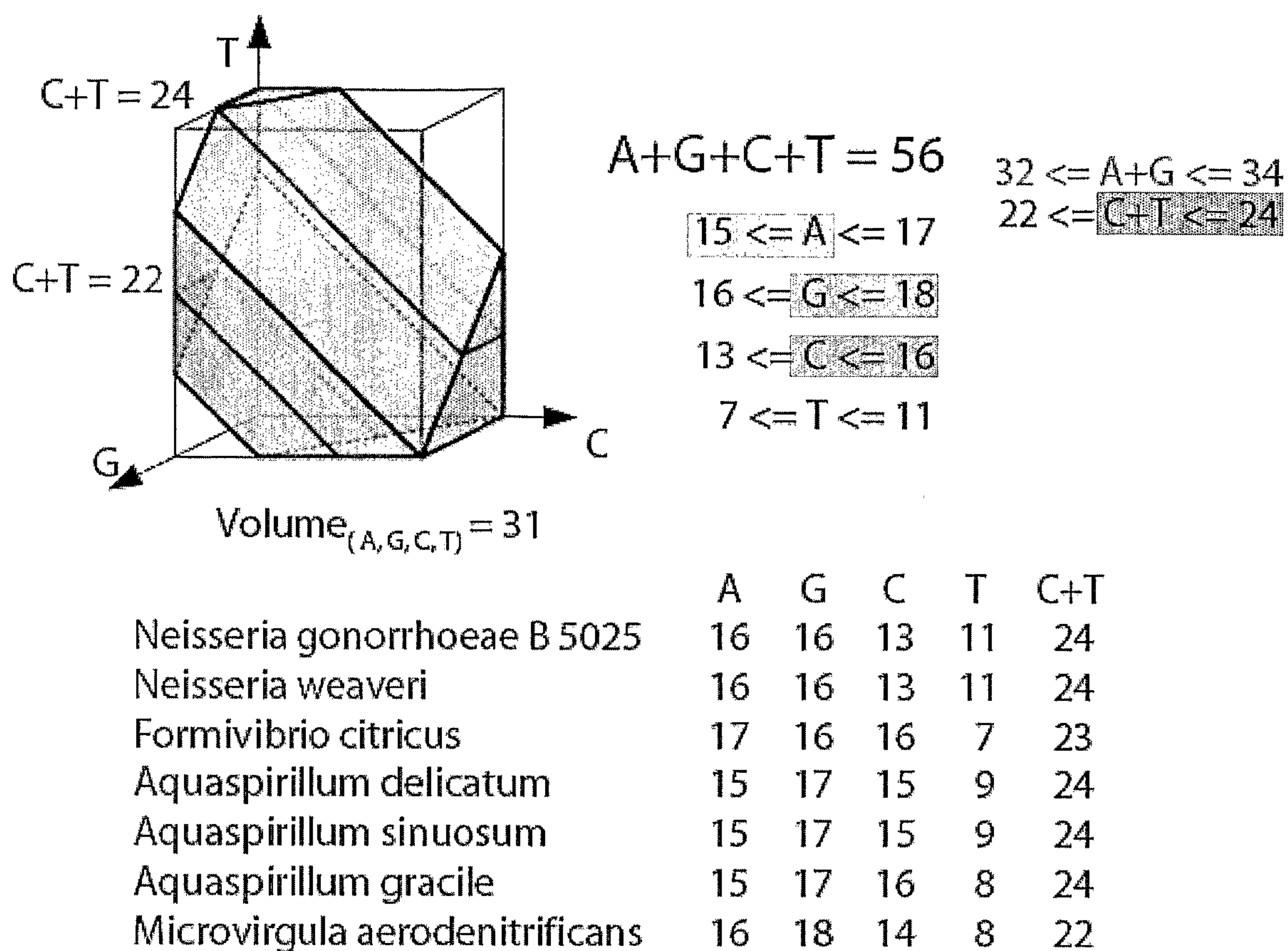


Figure 8A

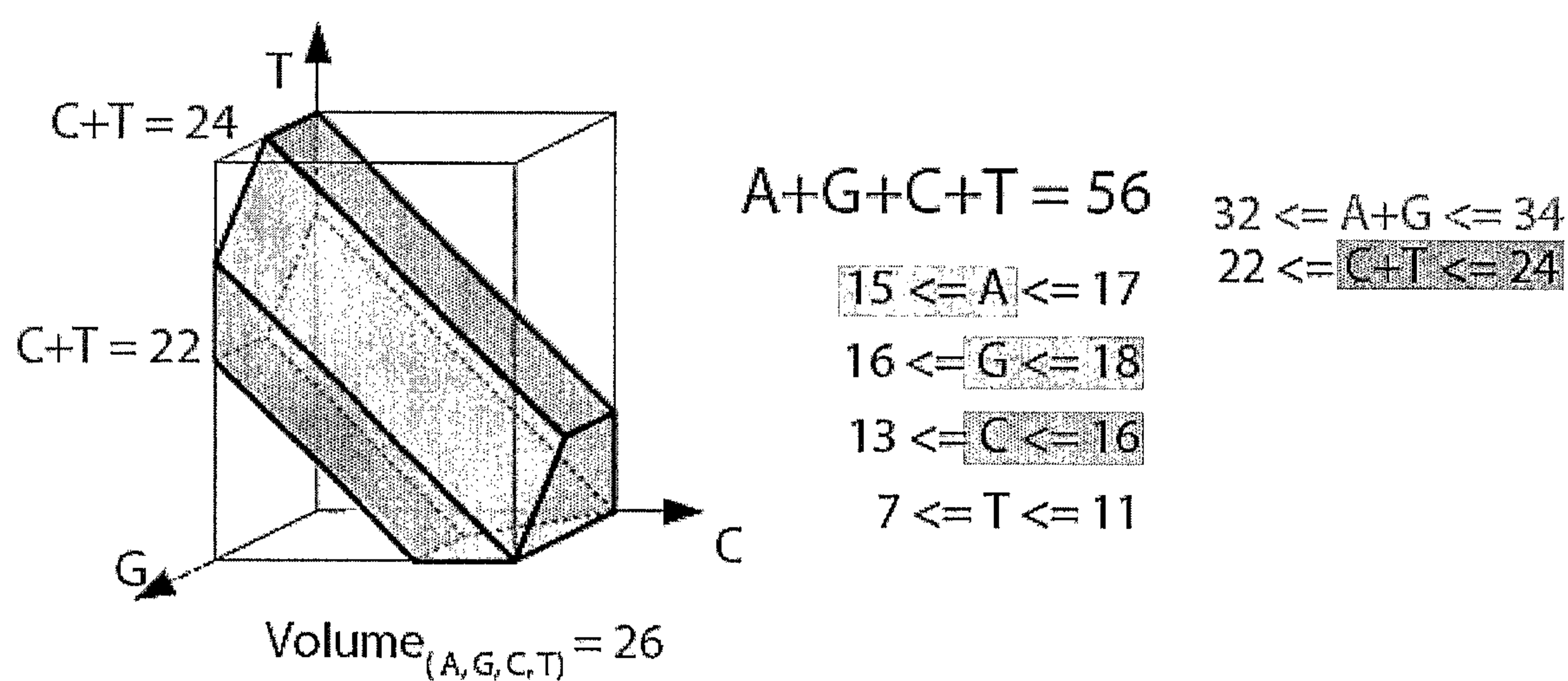
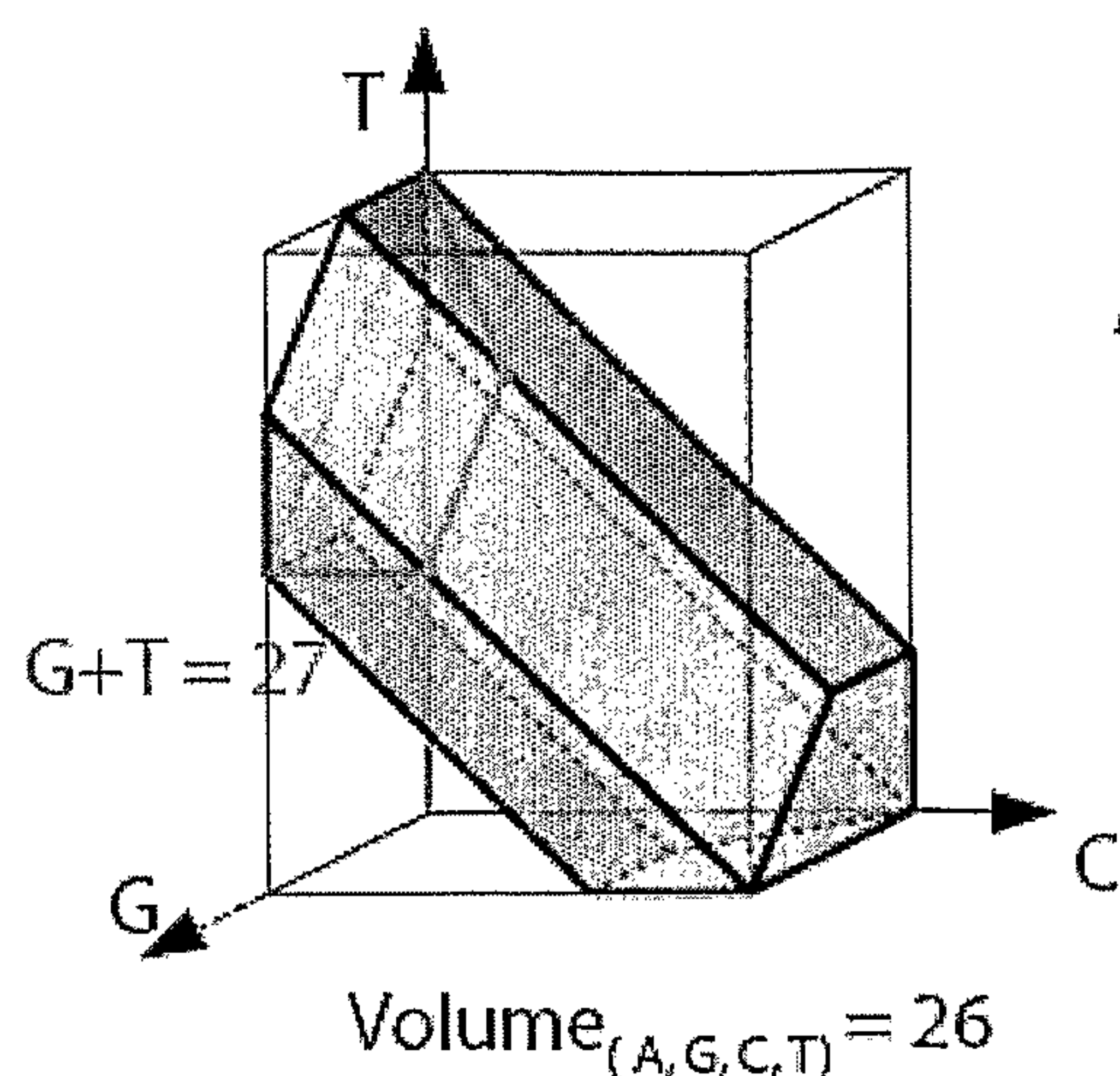


Figure 8B



|                                       | A  | G  | C  | T  | C+T | G+T |
|---------------------------------------|----|----|----|----|-----|-----|
| <i>Neisseria gonorrhoeae</i> B 5025   | 16 | 16 | 13 | 11 | 24  | 27  |
| <i>Neisseria weaveri</i>              | 16 | 16 | 13 | 11 | 24  | 27  |
| <i>Formivibrio citricus</i>           | 17 | 16 | 16 | 7  | 23  | 23  |
| <i>Aquaspirillum delicatum</i>        | 15 | 17 | 15 | 9  | 24  | 26  |
| <i>Aquaspirillum sinuosum</i>         | 15 | 17 | 15 | 9  | 24  | 26  |
| <i>Aquaspirillum gracile</i>          | 15 | 17 | 16 | 8  | 24  | 25  |
| <i>Microvirgula aerodenitrificans</i> | 16 | 18 | 14 | 8  | 22  | 26  |

Figure 9A

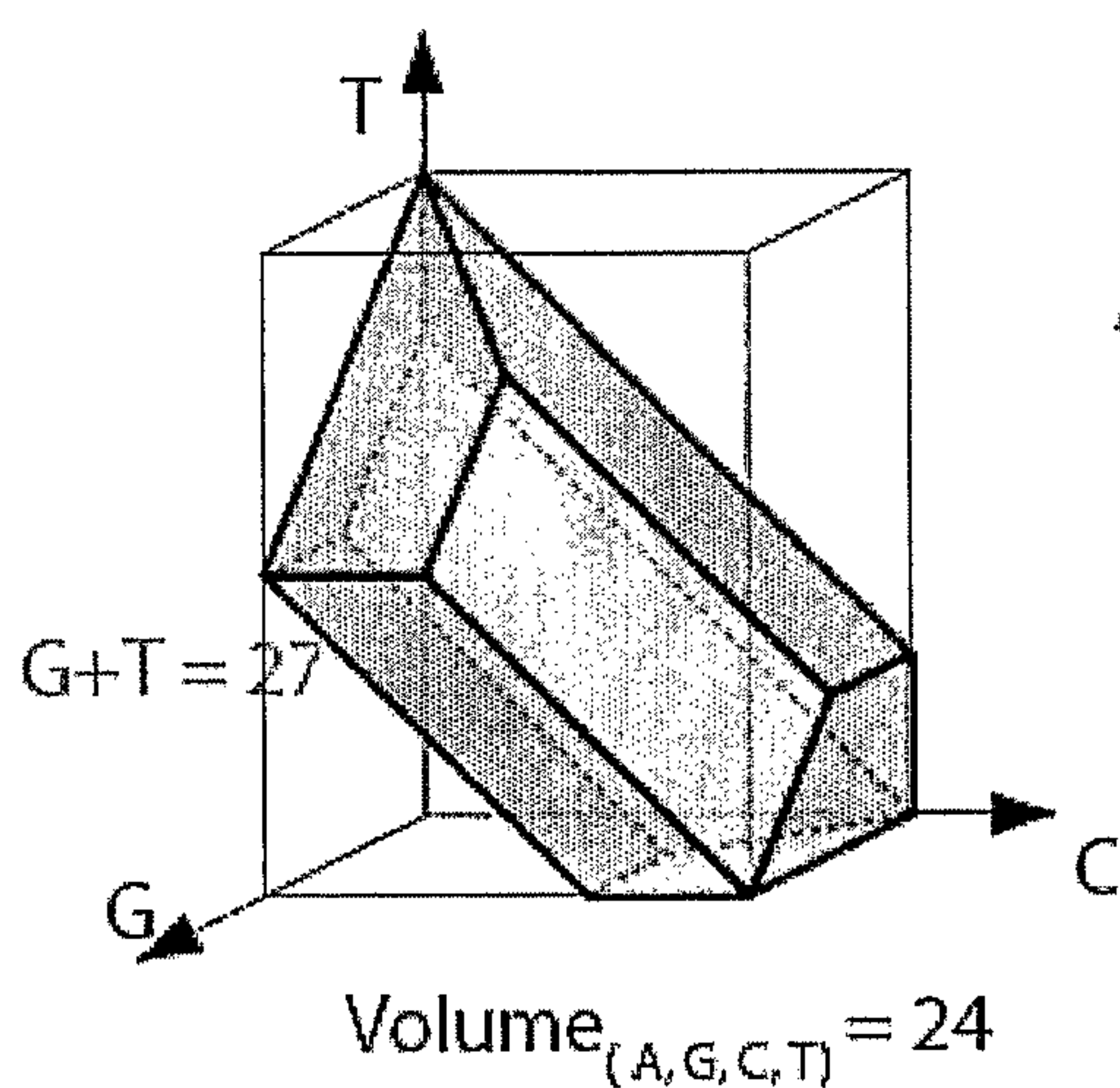
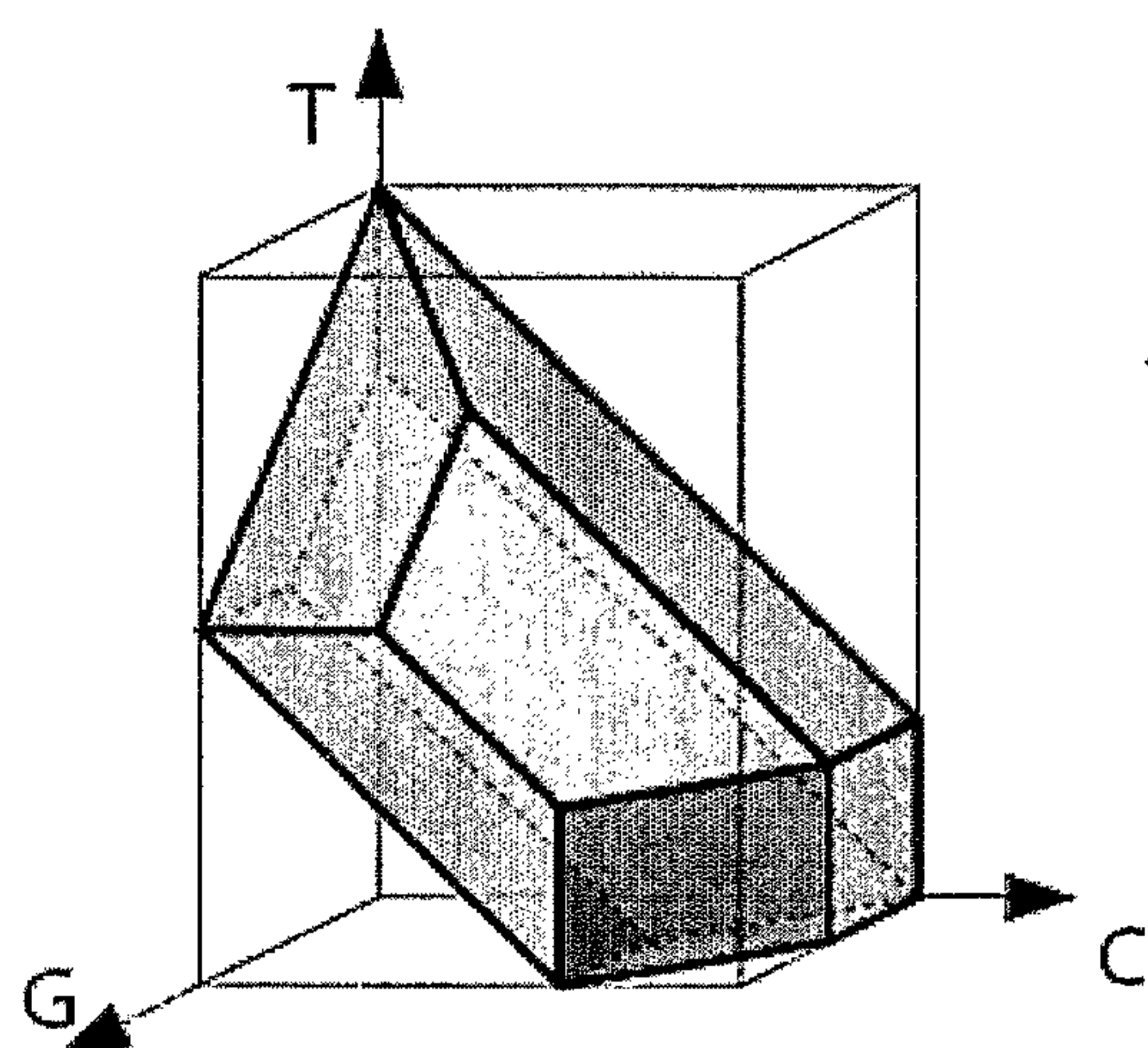


Figure 9B



$$A+G+C+T = 56$$

$$15 \leq A \leq 17$$

$$16 \leq G \leq 18$$

$$13 \leq C \leq 16$$

$$7 \leq T \leq 11$$

$$32 \leq A+G \leq 34$$

$$22 \leq C+T \leq 24$$

$$29 \leq A+C \leq 33$$

$$23 \leq G+T \leq 27$$

$$23 \leq A+T \leq 27$$

$$29 \leq G+C \leq 33$$

|                                       | A  | G  | C  | T  | C+T | G+T | G+C |
|---------------------------------------|----|----|----|----|-----|-----|-----|
| <i>Neisseria gonorrhoeae</i> B 5025   | 16 | 16 | 13 | 11 | 24  | 27  | 29  |
| <i>Neisseria weaveri</i>              | 16 | 16 | 13 | 11 | 24  | 27  | 29  |
| <i>Formivibrio citricus</i>           | 17 | 16 | 16 | 7  | 23  | 23  | 32  |
| <i>Aquaspirillum delicatum</i>        | 15 | 17 | 15 | 9  | 24  | 26  | 32  |
| <i>Aquaspirillum sinuosum</i>         | 15 | 17 | 15 | 9  | 24  | 26  | 32  |
| <i>Aquaspirillum gracile</i>          | 15 | 17 | 16 | 8  | 24  | 25  | 33  |
| <i>Microvirgula aerodenitrificans</i> | 16 | 18 | 14 | 8  | 22  | 26  | 32  |

Population = 7,  $\text{Volume}_{(A,G,C,T)} = 23 \Rightarrow \text{density} = 7/23 = 0.304$

Figure 10

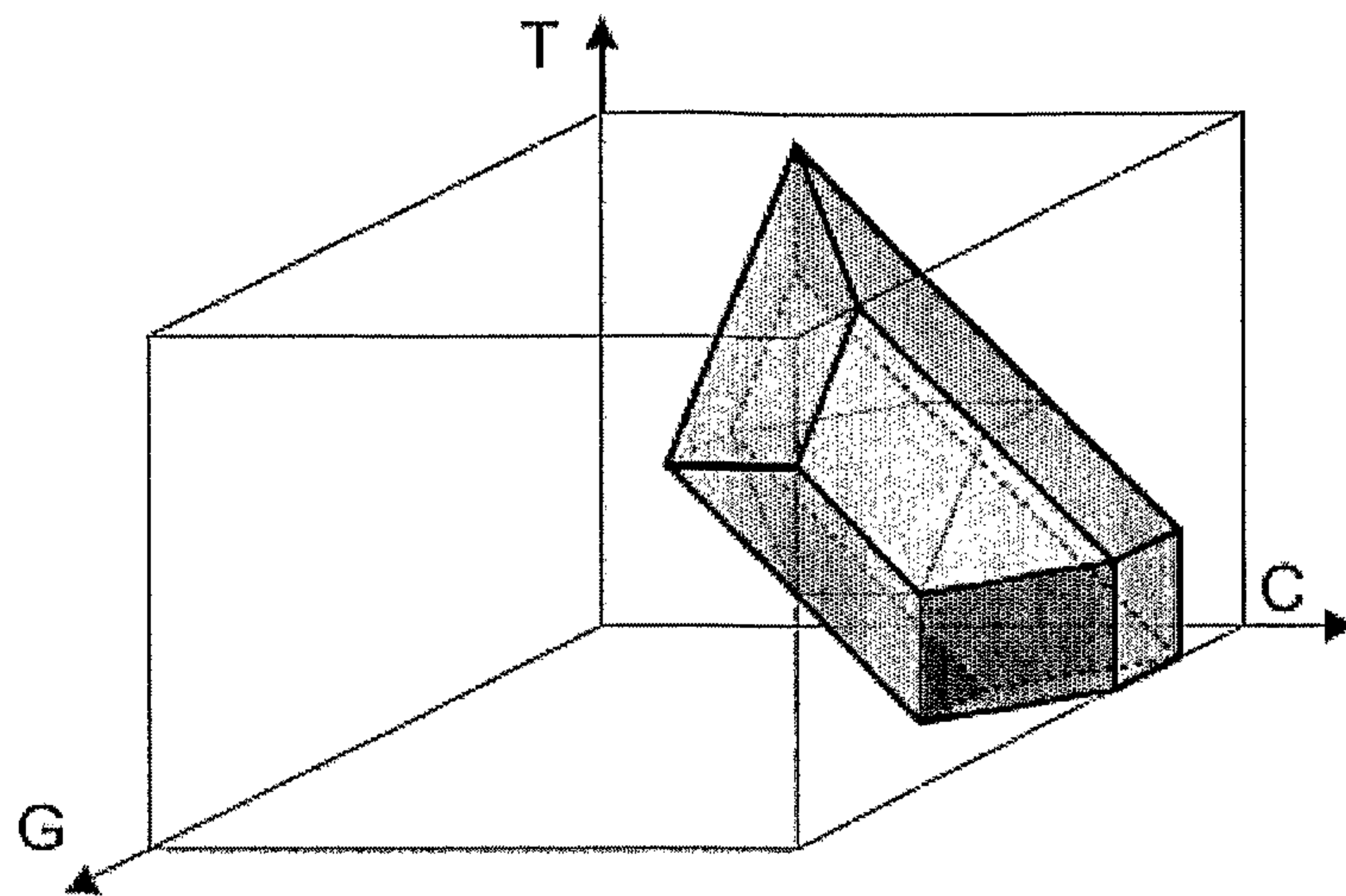


Figure 11A

| Taxon        | Pop. | Vol. | Density |
|--------------|------|------|---------|
| Neisseriales | 7    | 23   | 0.304   |

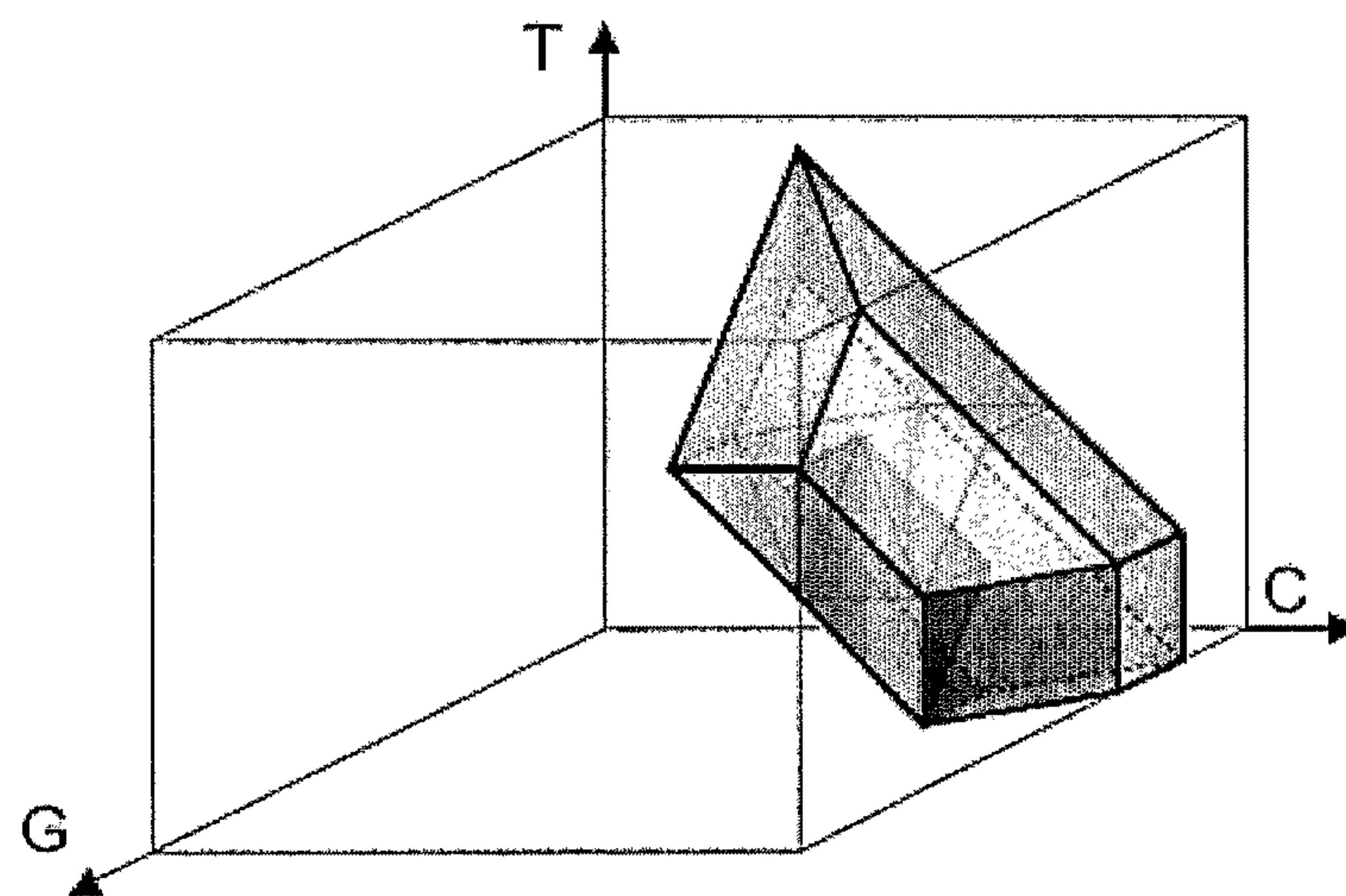


Figure 11B

| Taxon            | Pop. | Vol. | Density |
|------------------|------|------|---------|
| Neisseriales     | 7    | 23   | 0.304   |
| Nitrosomonadales | 8    | 6    | 1.333   |

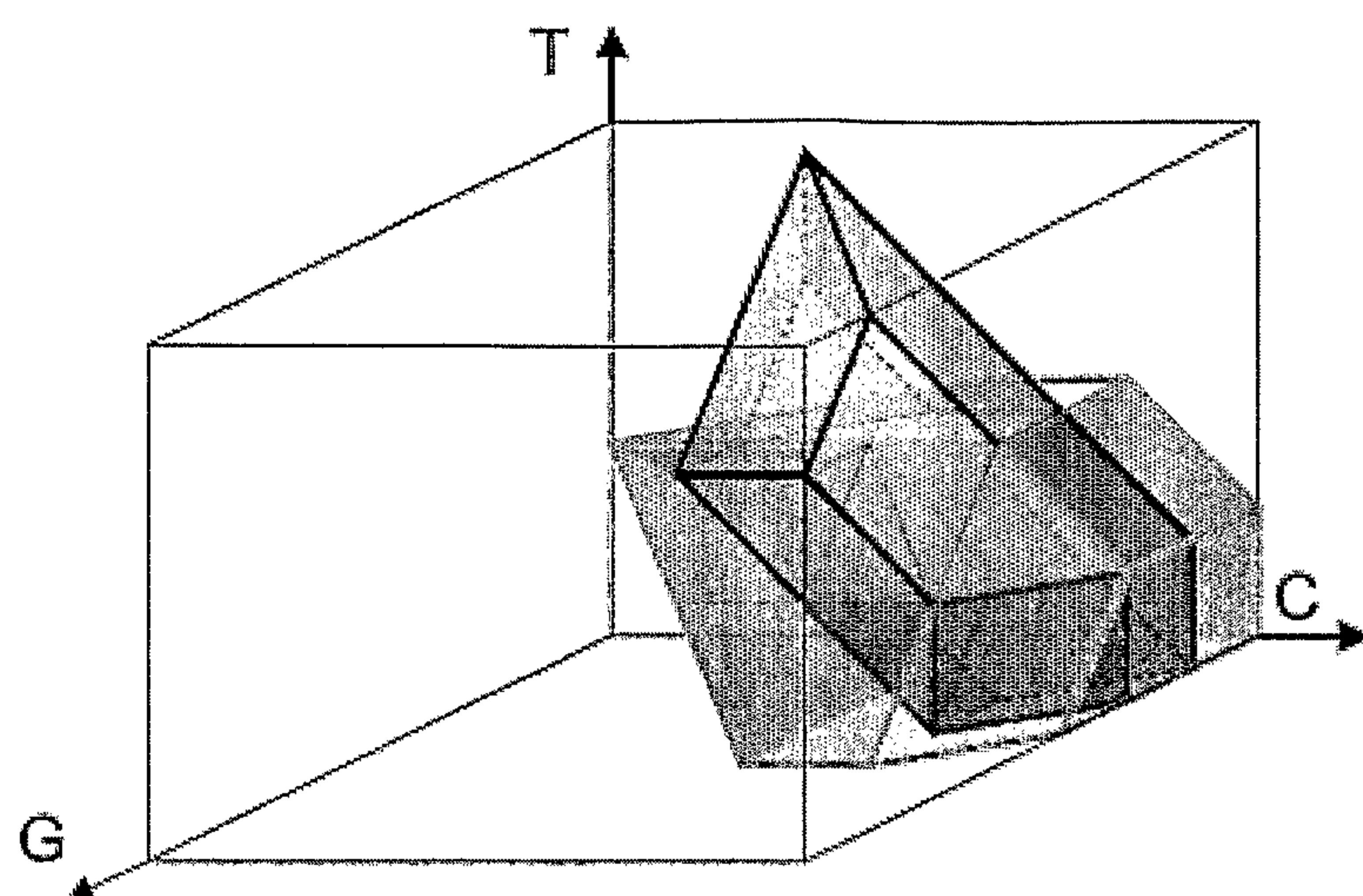


Figure 11C

| Taxon            | Pop. | Vol. | Density |
|------------------|------|------|---------|
| Neisseriales     | 7    | 23   | 0.304   |
| Nitrosomonadales | 8    | 6    | 1.333   |
| Burkholderiales  | 102  | 36   | 2.833   |

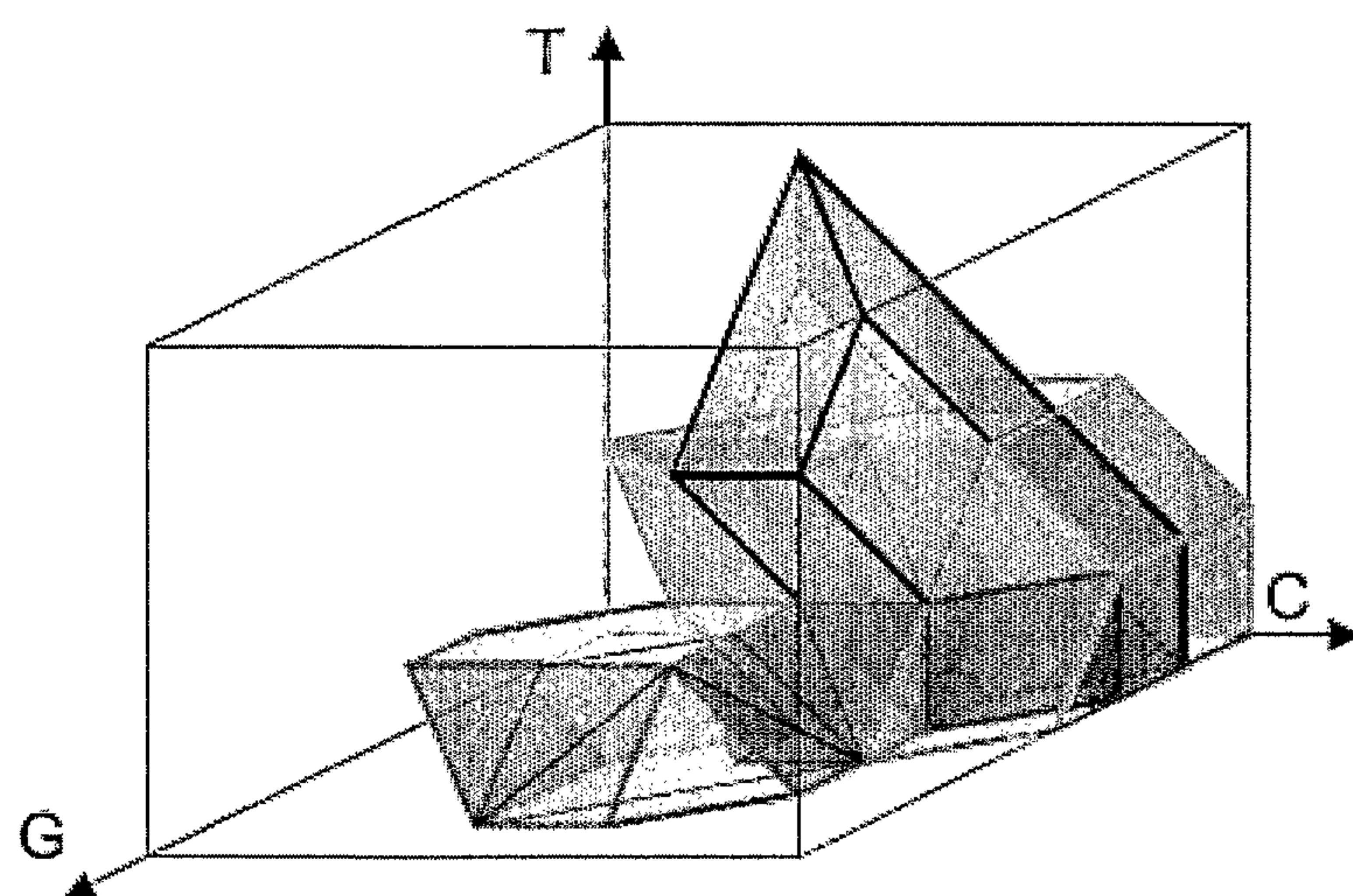


Figure 11D

| Taxon              | Pop. | Vol. | Density |
|--------------------|------|------|---------|
| Neisseriales       | 7    | 23   | 0.304   |
| Nitrosomonadales   | 8    | 6    | 1.333   |
| Burkholderiales    | 102  | 36   | 2.833   |
| Hydrogenophillales | 5    | 18   | 0.278   |

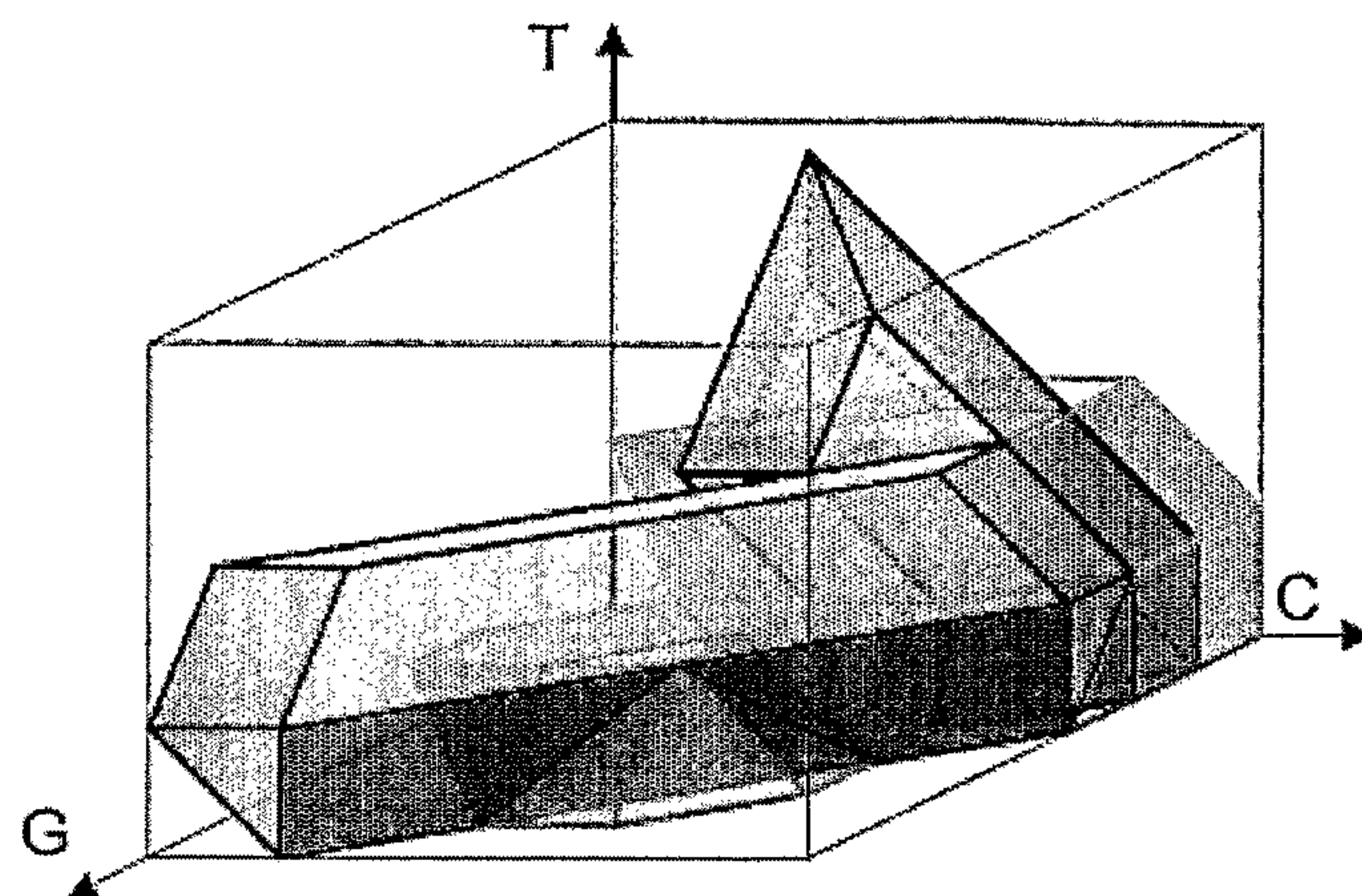


Figure 11E

| Taxon             | Pop. | Vol. | Density |
|-------------------|------|------|---------|
| Neisseriales      | 7    | 23   | 0.304   |
| Nitrosomonadales  | 8    | 6    | 1.333   |
| Burkholderiales   | 102  | 36   | 2.833   |
| Hydrogenophilales | 5    | 18   | 0.278   |
| Rhodocyclales     | 14   | 25   | 0.560   |

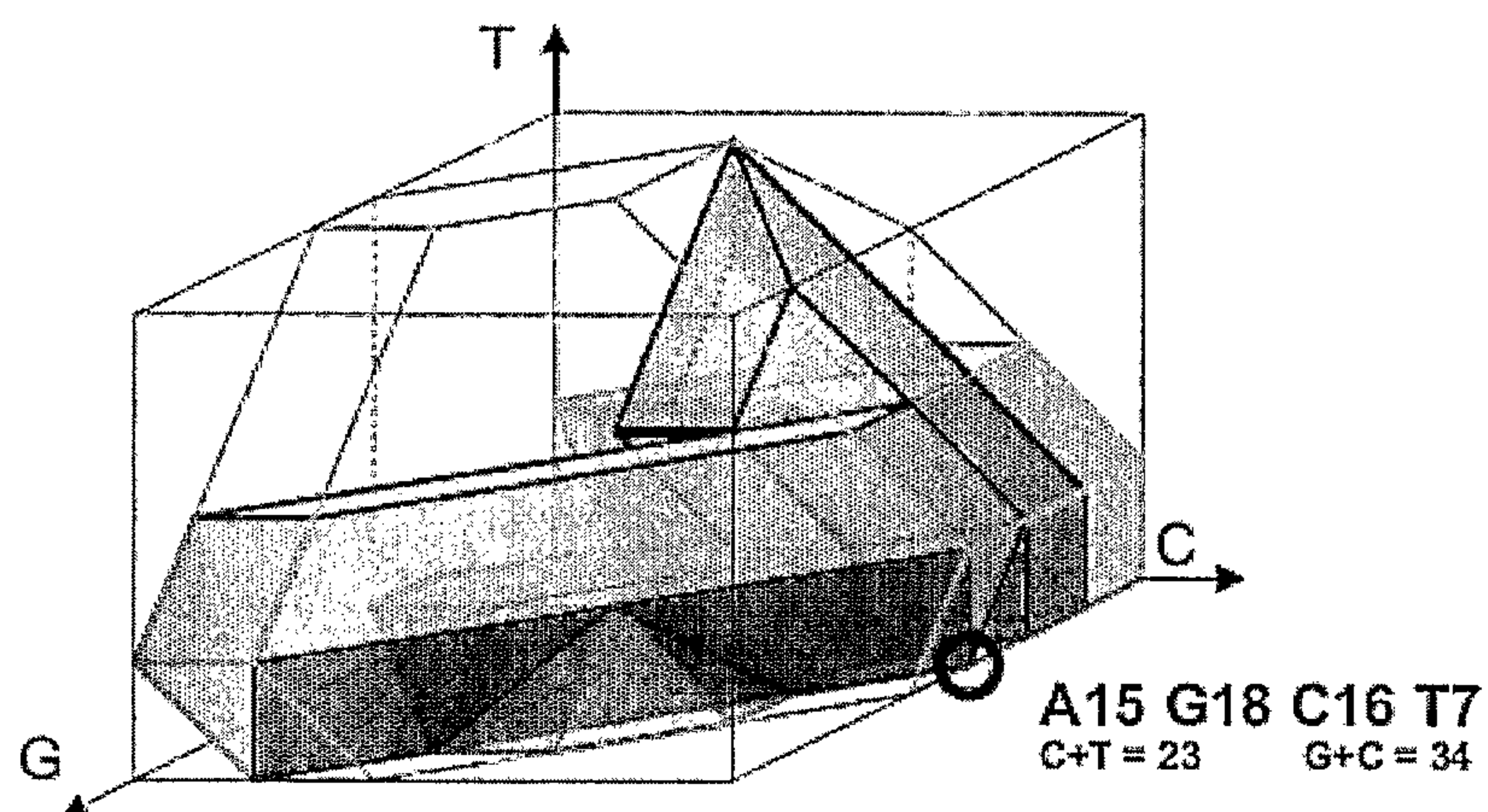


Figure 11F

| Polyhedron A+G+C+T=56     |      |      |         | Test base composition         |                  |                      |
|---------------------------|------|------|---------|-------------------------------|------------------|----------------------|
| Taxon                     | Pop. | Vol. | Density | min Distance<br>to Polyhedron | Point<br>Density | Match<br>probability |
| Neisseriales              | 7    | 23   | 0.304   | 1                             | 0.001            | 0.03%                |
| Nitrosomonadales          | 8    | 6    | 1.333   | 1                             | 0.005            | 0.15%                |
| Burkholderiales           | 102  | 36   | 2.833   | 0                             | 2.833            | 83.08%               |
| Hydrogenophilales         | 5    | 18   | 0.278   | 1                             | 0.001            | 0.03%                |
| Rhodocyclales             | 14   | 25   | 0.560   | 0                             | 0.560            | 16.42%               |
| Betaproteobacteria orders |      |      |         |                               | 3.401            | 99.71%               |

| Taxon             | Individual primer pair match probability |                      |                      |                      | Assignment probability |
|-------------------|------------------------------------------|----------------------|----------------------|----------------------|------------------------|
|                   | I                                        | II                   | III                  | IV                   |                        |
| Rhodocyclales     | 16.42%                                   | 45.89                | 22.80%               | 9.58%                | 96.769%                |
| Neisseriales      | 0.03%                                    | < 10 <sup>-6</sup>   | 1 x 10 <sup>-6</sup> | 6.40%                | 0.003%                 |
| Burkholderiales   | 83.08%                                   | 5 x 10 <sup>-6</sup> | 1 x 10 <sup>-6</sup> | 0.04%                | 0.071%                 |
| Hydrogenophilales | 0.03%                                    | 7 x 10 <sup>-6</sup> | 1 x 10 <sup>-6</sup> | < 10 <sup>-6</sup>   | 0.000%                 |
| Nitrosomonadales  | 0.15%                                    | < 10 <sup>-6</sup>   | < 10 <sup>-6</sup>   | 3 x 10 <sup>-6</sup> | 0.000%                 |

| Phylum                       | Class                        | Order                        | Family                       | Genus                        |
|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| Betaproteobacteria<br>82.36% | Betaproteobacteria<br>82.36% | Betaproteobacteria<br>82.36% | Betaproteobacteria<br>82.36% | Betaproteobacteria<br>82.36% |

Figure 12

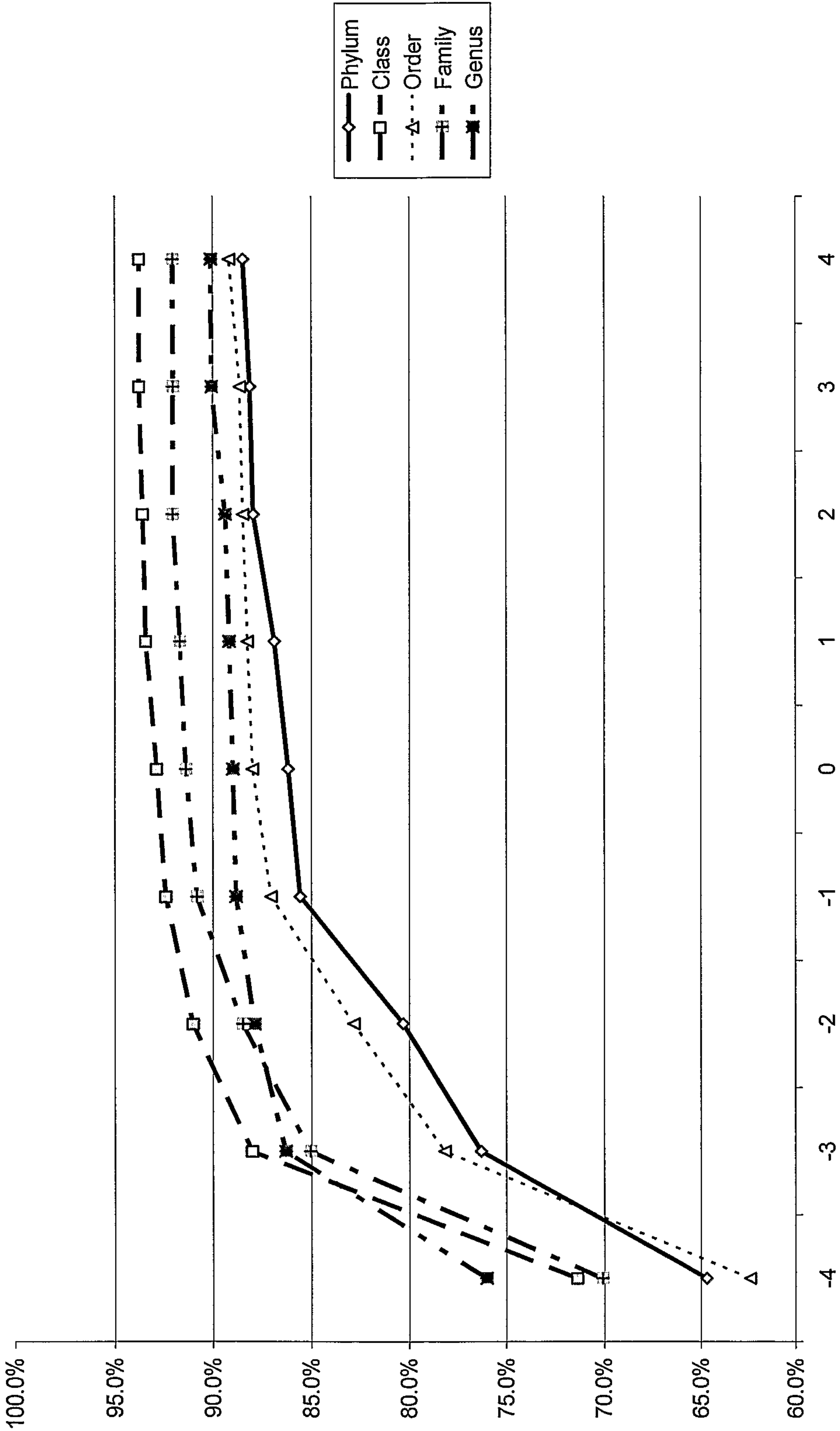


Figure 13

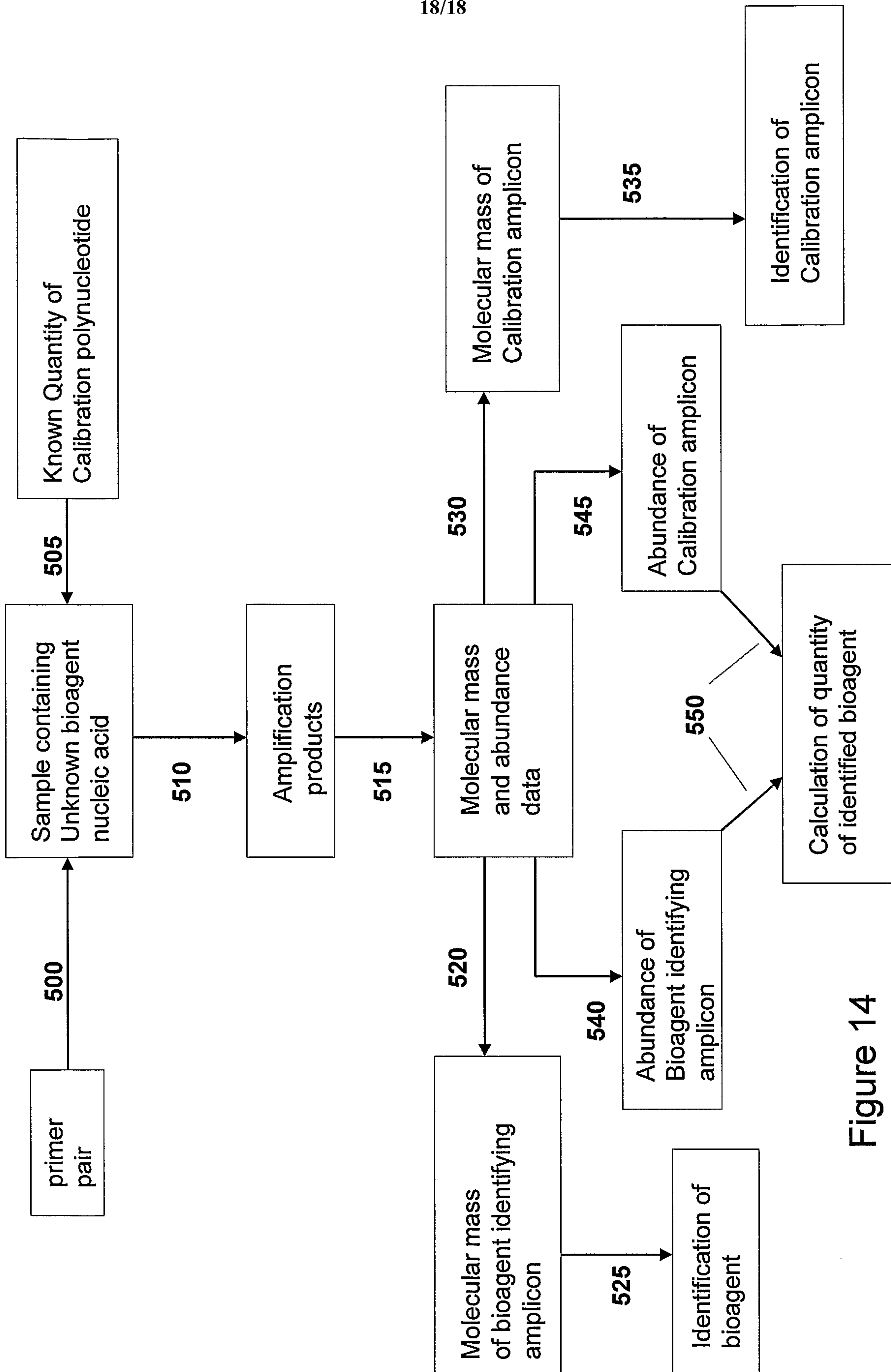


Figure 14