Title: DIAGNOSIS OF TUBERCULOSIS

Abstract: The invention provides a method of diagnosing tuberculosis (TB) in a test subject, said method comprising: (i) providing expression data of two or more markers in a subject, wherein at least two of said markers are selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-AI (Apo-AI), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp671032; and (ii) comparing said expression data to expression data of said marker from a group of control subjects, wherein said control subjects comprise patients suffering from inflammatory conditions other than TB, thereby determining whether or not said test subject has TB.
DIAGNOSIS OF TUBERCULOSIS

Field of the Invention

The present invention relates to the diagnosis of tuberculosis (TB).

Background of the Invention

Latent TB is present in one third of the world’s population with a prevalence of active TB in many geographic areas exceeding 700 cases per 100,000 of the population (WHO Stop TB www.who.int/grb). This global TB epidemic is fuelled through synergy with HIV, which is found in 40%-70% of African patients with active TB. In areas of high TB prevalence, sputum smear microscopy is often the only available and affordable test but at best achieves a sensitivity of 50%. Culture of Mycobacterium tuberculosis, the diagnostic gold standard, increases sensitivity by a further 25%. Tuberculin skin tests are often insufficiently accurate to aid diagnosis, particularly in areas of high TB prevalence. Serological tests for TB have focused on detection of mycobacterial antigen(s) and, like skin tests, are frequently confounded by cross-reactivity with non-pathogenic mycobacteria or previous immunisation with BCG.

Most deaths from tuberculosis (TB) are preventable by early diagnosis and treatment. Early diagnosis also minimises morbidity and risk of transmission and commonly relies on microscopic identification of Mycobacterium tuberculosis. However microscopy is insensitive and culture of organisms is often too slow to aid therapeutic decisions. Recently developed DNA amplification and interferon-gamma based tests are expensive and need particular expertise.

An accurate and rapid diagnostic test for TB will have immense impact on the control of this disease.

Summary of the Invention

The present inventors have applied supervised machine-learning analysis to proteomic profiles, and have successfully distinguished patients with active TB from control patients with overlapping clinical features. The inventors have achieved a diagnostic accuracy of 94% for patients with TB and this is unaffected by ethnicity.
or HIV status. After ranking the most informative peaks in the proteomic profiles by feature selection, four polypeptides, serum amyloid A protein, transthyretin apolipoprotein-A1 and serum albumin, were identified and quantitated by immunoassay. Two of these polypeptides, serum amyloid A and transthyretin, reflect inflammatory states, and so the inventors also quantitated neopterin and C reactive protein. In addition, apolipoprotein-A2, hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein and hypothetical protein DFKZp667I032 were identified as markers of TB by analysing the 2D gels used to identify peaks in the proteomic profile. Application of support vector machine classifiers to combinations of these markers gave a diagnostic accuracy of up to 84% for TB.

Accordingly, the present invention provides:

- a method of diagnosing tuberculosis (TB) in a test subject, said method comprising:
  
  (i) providing expression data of two or more markers in a test subject, wherein at least two of said markers are selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp667I032; and
  
  (ii) determining whether expression of said markers is indicative of TB by comparing said expression data to expression data of said two or more markers from a group of control subjects, wherein said group of control subjects comprises patients suffering from inflammatory conditions other than TB,

thereby determining whether or not said test subject has TB;

- a method of a method of diagnosing tuberculosis (TB), said method comprising:

  (i) providing expression data of two or more markers in a subject, wherein at least two of said markers are selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2),
hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL (LRG1)) and hypothetical protein DFKZp6671032; and

(ii) determining whether expression of said markers is indicative of TB;

5 - a method of diagnosing tuberculosis (TB), said method comprising:

(i) providing expression data of two or more markers in a subject, wherein at least two of said markers are selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp6671032; and

(ii) determining whether expression of said markers is indicative of TB, wherein said determination is implemented using a computer system programmed with a trained machine learning classifier;

10 - a computer-implemented method of diagnosing TB, said method comprising:

(i) inputting expression data of two or more markers in a subject; and

(ii) determining whether expression of said markers is indicative of TB using a computer system programmed with a trained support vector machine (SVM)

15 thereby diagnosing whether or not said patient has TB;

- a method of training a support vector machine (SVM) classifier to diagnose tuberculosis (TB), said method comprising:

(i) providing training data which comprises:

(a) training data relating to two or more markers in each of a first set of TB patients; and

(b) training data relating to said two or more markers in each of a first set of control subjects;

(ii) using a SVM to discriminate the training data of TB patients from the training data of control subjects;

25 thereby training the SVM to diagnose TB;

- an apparatus arranged to perform a method according to the invention comprising:
(i) means for receiving expression data of two or more markers in a sample from a subject;
(ii) a module for determining whether said data is indicative of TB, wherein said module comprises a trained machine learning classifier capable of distinguishing data from a TB patient from data from a control subject; and
(iii) means for indicating the results of said determination;
- a computer program executable by a computer system, the computer program being capable, on execution by the computer system, of causing the computer system to perform a method according to the invention;
- a storage medium storing in a form readable by a computer system having a computer program according to the invention;
- a kit for diagnosing TB comprising:
  (i) means for detecting two or more markers; and
  (ii) a storage medium according to the invention;
- a kit for diagnosing TB comprising:
  (i) means for detecting two or more markers;
  (ii) instructions for inputting data relating to detection of said markers into an apparatus according to the invention;
- a kit for diagnosing TB comprising:
  (i) means for detecting two or more markers selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp667I032;
- a method of identifying an agent for the treatment of TB, said method comprising:
  (i) contacting a test agent with a TB marker selected from transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2, hemoglobin beta, haptoglobin, DEP domain protein and A2GL; and
  (ii) determining whether said test agent modulates the activity or expression of said marker,
thereby determining whether or not said test agent is suitable for use in the treatment of TB; and
- a method of identifying an agent for the treatment of TB, said method comprising:
  (i) contacting cells ex vivo or in vivo with *Mycobacterium tuberculosis* and a test agent;
  (i) monitoring expression of one or more TB markers selected from transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2, hemoglobin beta, haptoglobin, DEP domain protein and A2GL; and
(ii) determining whether test agent modulates the expression of said one or more test markers,
thereby determining whether or not said test agent is suitable for use in the treatment of TB.

15 **Brief Description of the Figures**

Figure 1 is a flow chart of a method of training a machine learning classifier.

Figure 2 is a flow chart of a method of testing a trained machine learning classifier.

Figure 3 is a flow chart of a method of determining whether a subject has or does not have TB using a trained machine learning classifier.

Figure 4 shows the parameterisation of Gaussian kernel sigma value of Classifier (SVM_1 in Table 3). The Gaussian SVM was trained with the initial training set (Table 2) using all mass peak clusters (10-fold cross validation for parameter selection). Classifier performance was then assessed on the initial testing set (Table 2).

Figure 5 shows the averaged ROC using 10-fold train cross validation test. One hundred randomly selected train and test sets with a train:test ratio (80:20) were created. Parameters were selected using a 10-fold cross validation on the train set and performance obtained in the corresponding test set. a) Upper line shows the averaged ROC curve of the classifiers obtained when kernel parameter is selected on sensitivity criteria. b) Upper line shows the averaged ROC curve of the classifiers obtained when kernel parameters is selected on specificity criteria.
Brief Description of the Sequences

SEQ ID NO: 1 is the amino acid sequence of human serum amyloid A1.
SEQ ID NO: 2 is the amino acid sequence of human C-reactive protein.
SEQ ID NO: 3 is the amino acid sequence of human transthyretin.
SEQ ID NO: 4 is the amino acid sequence of human serum albumin precursor.
SEQ ID NO: 5 is the amino acid sequence of human apolipoprotein-A1.
SEQ ID NO: 6 is the amino acid sequence of human leucine-rich alpha-2-glycoprotein.
SEQ ID NO: 7 is the amino acid sequence of human hemoglobin beta.
SEQ ID NO: 8 is the amino acid sequence of human haptoglobin.
SEQ ID NO: 9 is the amino acid sequence of human apolipoprotein-A2.
SEQ ID NO: 10 is the amino acid sequence of human DEP domain protein.
SEQ ID NO: 11 is the amino acid sequence of human hypothetical protein DFKZp667I032.

Detailed Description of the Invention

The present invention provides an ex vivo method of diagnosing tuberculosis (TB) in a test subject, said method comprising or consisting essentially of the steps of:

(i) providing expression data of two or more markers in a test subject, wherein at least two of said markers are selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp667I032; and
(ii) determining whether expression of said markers is indicative of TB by comparing said expression data to expression data of said marker from a group of control subjects, wherein said group of control subjects comprises patients suffering from inflammatory conditions other than TB, thereby determining whether or not said test subject has TB.
The group of control subjects may be selected from one or more patients with respiratory infections, patients with sarcoidosis, patients with inflammatory bowel disease, patients with malaria, patients with human African trypanosomiasis (HAT), patients with neurological disease, patients with autoimmune disease, patients with myeloma and healthy subjects.

The present invention provides an *ex vivo* method of diagnosing tuberculosis (TB), said method comprising or consisting essentially of the steps of:

(i) providing expression data of two or more markers in a subject, wherein at least two of said markers are selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp667I032; and

(ii) determining whether expression of said markers is indicative of TB, thereby diagnosing whether or not patient has TB.

A marker is a molecule, such as a protein or peptide, which is differentially expressed in a sample taken from a TB patient as compared to an equivalent sample or samples taken from one or more control subjects who do not have TB. The expression data typically provides an indication of the amount of marker present in a sample from a subject. A marker is present differentially in samples taken from TB patients and samples taken from control subjects if it is present at an increased level (positive marker) or a decreased level (negative marker) in TB samples compared to control samples. Preferably, the increase or decrease in the amount of a marker is a statistically significant difference.

The term 'sensitivity' is herein defined as the conditional probability of a true positive. The term 'specificity' is herein defined as the conditional probability of a true negative. The term 'accuracy' is herein defined as the proportion of correct classifications. Hence, accuracy indicates the reproducibility of the specific marker pairs or clusters for diagnosis of TB; sensitivity indicates how likely the combination was of achieving a true positive diagnosis; and specificity indicated how well each marker combination was in identifying samples as a true negative for TB infection.
Transthyretin, neopterin, CRP and SAA are known to be associated with pathophysiological processes in TB. However, it has not previously been suggested that any of these proteins may be used as markers in the diagnosis of TB. The present inventors have identified SAA, neopterin, CRP, serum albumin, Apo-A1, A2GL and DEP domain protein as positive markers of TB and transthyretin, Apo-A2, hemoglobin beta, haptoglobin and hypothetical protein DFKZp6671032 as negative markers of TB. The present inventors have found that when used in various combinations, these markers, and in particular SAA, neopterin, CRP and transthyretin, can be used to diagnose TB with a high degree of sensitivity, specificity and accuracy. Methods of the invention typically allow diagnosis of TB with an accuracy, a specificity and/or a sensitivity of at least 80%, for example, at least 85%, at least 90% or at least 95%.

The present invention thus allows determination of whether a subject is infected with *Mycobacterium tuberculosis* quickly and easily without the need to culture *Mycobacterium tuberculosis* in a sample from said subject. The method of the present invention enables TB to be distinguished from other infections such as viral and bacterial infectious and inflammatory diseases other than TB. Examples of infections and inflammatory diseases that may be distinguished from TB include other respiratory infections, sarcoidosis, inflammatory bowel disease, malaria, human African trypanosomiasis, neurological disease, autoimmune disease and myeloma.

In a method of the invention the expression data from the subject is typically compared to expression data of the same markers in a TB patient. The TB patient may have been diagnosed as having TB by culture of *Mycobacterium tuberculosis* from a sample from the patient. The expression data may also be compared to expression data of the same marker in one or more control subject. The control subject may be a patient having an inflammatory disease other than TB. The inflammatory disease may be caused by a pathogenic infection, for example a bacterial, viral or fungal infection. The control subject may have any of the diseases other than TB mentioned herein. Alternatively or additionally, one or more of the control subjects may be healthy individuals. A healthy individual is an individual not having an inflammatory disease.
Use of expression data from two or more markers enhances the accuracy of the diagnosis. Using combinations of more than two markers, such as three or more markers, may further enhance the accuracy of diagnosis. Accordingly, expression data from two or more markers, preferably three or more markers, for example four or more markers, such as five, six, seven, eight, nine, ten, fifteen, twenty or more markers, is used in a method of the invention. It is preferable that one of these markers used in the method of diagnosis is transthyretin. Preferred combinations include (i) transthyretin, SAA and CRP, (ii) transthyretin and neopterin and (iii) transthyretin, neopterin and CRP. Additional markers, such as serum albumin and/or Apo-A1, other than transthyretin, neopterin, SAA and CRP may be included in the analysis. Further additional markers include apolipoprotein-A2, hemoglobin beta, haptoglobin protein, DEP domain protein, A2GL and hypothetical protein DFKZp6671032.

Further additional markers may be proteins or peptides that are present at elevated or reduced levels in TB samples compared to control samples. The additional marker(s) may be characterised by an apparent molecular weight or mass-to-charge ratio (m/z value), for example as determined by mass spectrometry.

Such additional biomarkers may be identified by the method used by the present inventors to determine that SAA, serum albumin and Apo-A1 are positive markers of TB and that transthyretin is a negative marker of TB. Other positively and negatively correlated markers may be identified by surface enhanced laser desorption and ionization (SELDI) technology and supervised machine learning classification methods.

For example, the present inventors have identified ten positive markers and ten negative markers by comparing the proteomic signatures from TB patients with proteomic signatures from control subjects using a support vector machine classifier. The positive markers have m/z values of about M18394_9, about M8952_75, about M11720_0, about M11454_1, about M18591_2, about M11488_1, about M9076_68, about M8895_13, M10856_8 and about M11541_5 and the negatively correlated markers have m/z values of about M4100_03, about M3898_52, about M13972_1, about M3322_01, about M2956_45, about M5644_96, about M3939_63, about M4056_39, about M6649_74 and about M13774_3. The marker having an m/z value
of about M11541_5 is SAA. The marker having an \textit{m/z} value of about M18394_9 is serum albumin. The marker having an \textit{m/z} value of about M11454_1 is Apo-A1. The marker having an \textit{m/z} value of about M13774_3 is transthyretin. There may be some variation in \textit{m/z} value. For example, there may be variation that is dependent on the resolution of the machine used to determine \textit{m/z} value or on post-translational modification of the marker. Accordingly, the markers listed above may have the specified \textit{m/z} value plus or minus about 10\%, about 5\%, about 1\%, about 0.5\% or about 0.2\%.

The identity of the additional markers identified by SELDI analysis may be determined by tryptic digestion and Matrix-assisted laser desorption/ionization time of flight (MALDI-ToF) mass spectroscopy of the peptide mass fingerprints and comparison with protein databases such as the MASCOT database. SAA1 has an \textit{m/z} value of M11541_5 and transthyretin has an \textit{m/z} value of M13774_3 and were identified by such methods.

The markers may also be identified by identifying the protein spots corresponding to the \textit{m/z} value on a 2-dimensional (2D) gel and excising and identifying the protein present in the spot. The 2D gel may be obtained from pooled sera from a number, such as about 10, about 20 or more, of TB patients or a number, such as about 10, about 20 or more, of control subjects. The \textit{m/z} value is generally slightly smaller than the passive elution (PE) mass. The increase in the PE mass over the \textit{m/z} value is proportional to the time used to do the passive elution. Therefore, if this method is used it is important to note that the link between the \textit{m/z} value and the PE mass is approximate. However, the identity of the marker may be confirmed by immunodepleting the original sample and repeating the SELDI-ToF analysis. A reduction in the size of the peak with the \textit{m/z} value of interest indicates that a correct identification has been made. However, further identification is not essential for the proteins to be mass used as markers in a method of the invention. The positive markers having \textit{m/z} values of M18394_9 and M11454_1 have been identified as serum albumin precursor and apolipoprotein A1 (Apo-A1) using this method. Thus one or more of the markers identified by their \textit{m/z} values, including serum albumin and/or Apo1-A1, may be used as markers in a method of the invention.
Additional markers of TB may have been identified by identifying polypeptides that are differentially present in 2D gels containing serum proteins from TB patients and control subjects. The markers identified in this way are apolipoprotein A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein and hypothetical protein (DFKZp667j032) and leucine-rich-alpha-2-glycoprotein (A2GL (LRG1)).

Following supervised machine learning analysis of proteomic signatures from TB patients and control subjects, the protein clusters suitable for use as markers of TB may be identified by any method which enables selection of protein clusters with the power to discriminate between TB patients and control subjects. Typically, a correlation filter method is used to detect independently informative peaks. For example, the Pearson correlation coefficient may be used to rank peaks for their discriminatory power. The Pearson correlation coefficient is defined as

\[ R(k) = \frac{\text{covariance}(X_k, Y)}{\sqrt{\text{variance}(X_k) \text{variance}(Y)}} \]

where \( X_k \) is the random variable corresponding to the \( k^{th} \) component of sample input vectors \( x \) and \( Y \) is the random variable of output labels.

The estimate of \( R(k) \) is given by \( \hat{R}(k) = \frac{\sum_{i=1}^{m} (x_{ik} - \bar{x}_k)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_{ik} - \bar{x}_k)^2 \sum_{i=1}^{m} (y_i - \bar{y})^2}} \)

where \( x_{ik} \) correspond to value \( m/z \) of the mass cluster \( k \) of sample \( i \), \( y_i \) is the class label for sample \( i \) and \( m \) is the number of samples. \( R(i) \) may be used a test statistic to assess the significance of a variable and it is linked to the \( t \)-test. \( \hat{R}(k) \) may be calculated between values of each mass cluster and corresponding class labels across the training set. \( \hat{R}(k) \) may then be used to rank positively and negatively correlated mass clusters. Mass clusters with the highest positive and/or highest negative correlation coefficients may be selected.

Proteins are often present in biological material in a plurality of different forms characterised by detectably different molecular masses. Hence, analysis of expressed proteins in a biological sample by methods such as SELDI detects the various different forms of the protein as a protein cluster. The different forms may
result from pre-translational and/or post-translational modifications. For example, the transthyretin marker may be transthyretin precursor or mature transthyretin. As additional Examples, each of the serum albumin, Apo-A1 and Apo-A2 markers may also be a precursor or mature form of the protein, preferably a precursor form.

Allelic variation, the generation of splice variants and RNA editing give rise to pre-translational modifications. Post-translational modifications include proteolytic cleavage, glycosylation, phosphorylation, lipidation, oxidation, methylation, cystinylation, sulphonation and acetylation. The expression data may relate to any one or more form of the protein. Pre- and/or post-translational modifications may give rise to fluctuations in the $m/z$ value of a marker in SELDI-ToF.

In one embodiment of the invention, the expression data may relate to one or more peptide derived from the said markers. For example, the expression data of SAA may relate to expression of a peptide resulting from loss of the N-terminal arginine of SAA. The full sequence of SAA1 is shown in SEQ ID NO: 1.

The expression data may, in one embodiment, relate to a particular form of the marker. For example, the positive markers Apo-A1 may be the form having a molecular mass of about 11400 to about 11600 and/or the positive marker serum albumin may be the form having a molecular weight of about 18300 to about 18500 daltons (Da).

Expression data may be obtained by any suitable method. In one embodiment, the expression data indicates the presence or absence of each marker of interest. The expression data preferably provides an indication of the amount of each marker present in a sample from a subject, i.e. the data is quantitative. The expression data may additionally qualify the form of each marker, for example the form of the protein present.

Typically, expression data is obtained by capture of the markers on a solid phase, or surface, and detection of the captured markers. The surface is designed to select marker proteins from samples according to a general property of the markers being used or according to specific properties of the different protein markers. The surface is typically a bead, plate, membrane or chip on which one or more capture reagent is bound. The capture reagent may be a specific chromatographic surface. The chromatographic surface may be chemically or biochemically treated.
Chemically treated surfaces may be anionic, cationic, hydrophobic, hydrophilic or metal. Such chemically treated surfaces are capable of capturing proteins with a particular chemical property. Such chemically treated surfaces may comprise, for example, ion exchange materials, metal chelators, such as nitriloacetic acid or iminodiacetic acid, immobilised metal chelates, hydrophobic interaction adsorbents, hydrophilic interaction adsorbents, dyes, simple biomolecules, such as nucleotides, amino acids, simple sugars and fatty acids, and mixed mode adsorbents, such as hydrophobic attraction/electrostatic repulsion adsorbents.

In an embodiment where the surface is biochemically treated, the capture reagent is typically a specific binding reagent for a particular marker. In this embodiment, the surface typically comprises a specific binding reagent for each marker being used. A protein “specifically binds” to a marker when it binds with preferential or high affinity to the marker for which it is specific but does not bind, does not substantially bind or binds with only low affinity to other substances. The specific binding capability of a protein may be determined by any suitable method. A variety of protocols for competitive binding are well known in the art (see, for example, Maddox et al. (1993)).

The specific binding agent may be an antibody or antibody fragment specific for the marker. Suitable antibodies are available in the art. Antibodies and antibody fragments may also be generated using standard procedures known in the art.

The antibody may be a monoclonal or polyclonal antibody. Monoclonal antibodies are preferred. The binding proteins may also be, or comprise, an affinity ligand or an antibody fragment, which fragment is capable of binding to the marker. Such antibody fragments include Fv, F(ab’') and F(ab’’)2 fragments as well as single chain antibodies. Aptamers, antibodies and interacting fusion proteins may also be used as specific binding agents. The specific binding agent may recognize one or more form of the marker of interest.

Other biochemically treated surfaces may be coated with a nucleic acid molecule, such as a polypeptide, a polysaccharide, a lipid, a steroid or a conjugate molecule, such as a glycoprotein, a lipoprotein, a glycolipid or a nucleic acid (e.g. DNA)-protein conjugate.
Methods for coupling specific binding agents such as antibodies to a surface are well known in the art.

The surface may be a protein chip array. A protein chip array comprises discrete spots, typically of a diameter of 2mm, of capture reagents. The capture reagents at each spot on the array may be the same or different. Protein chip arrays suitable for use in the invention are well known in the art. For example, suitable chips are available from Ciphergen Biosystems and include CM10, IMAC-3, CM16, SAX2, H4, NP20, H50, Q-10, WCX-2, IMAC-30, LSAX-30, LWCX-30, IMAC-40, PS10, PS-20 and PG-20 protein chip arrays.

These protein biochips typically comprise an aluminium substrate in the form of a strip. The surface of the strip is coated with silicon dioxide. In the case of the NP-20 biochip, silicon oxide functions as a hydrophilic adsorbent to capture hydrophilic proteins. H4, H50, SAX-2, Q-10, WCX-2, CM-10, IMAC-3, IMAC-30, PS-10 and PS-20 biochips further comprise a functionalised, cross-linked polymer in the form of a hydrogel physically attached to the surface of the biochip or covalently attached through a silane to the surface of the biochip. The H4 biochip has isopropyl functionalities for hydrophilic binding. The H50 biochip has nonylphenoxypoly(ethylene glycol) methacrylate for hydrophobic binding. The SAX-2 and Q-10 biochips have quaternary ammonium functionalities for anion exchange. The WCX-2 and CM-10 biochips have carboxylate functionalities for cation exchange. The IMAC-3 and IMAC-30 biochips have nitriloacetic acid functionalities that adsorb transition metal ions, such as Cu\(^{2+}\) and Ni\(^{2+}\), by chelation. These immobilised metal ions allow adsorption of peptide and proteins by coordinate bonding. The PS-10 biochip has carboimidazole functional groups that can react with groups on proteins for covalent binding. The PS-20 biochip has epoxide functional groups for covalent binding with proteins. The PS-series biochips are useful for binding biospecific adsorbents, such as antibodies, receptors, lectins, heparin, Protein A, biotin/streptavidin and the like, to chip surfaces where they function to specifically capture analytes from a sample. The PG-20 biochip is a PS-20 chip to which Protein G is attached. The LSAX-30 (anion exchange), LWCX-30 (cation exchange) and IMAC-40 (metal chelate) biochips have functionalised latex beads on their surfaces.
The surface may be a well of a microtitre plate, such as a 96-well microtitre plate. Typically, each well of such a plate will comprise a different capture reagent, such as a different antibody, as each well may comprise two or more discrete spots of different antibodies.

The capture surface may be a column loaded with a plurality of beads coated with the capture reagent. Multiple columns, each able to capture a single marker protein may be used. Alternatively, a single column may contain beads coated with specific binding agents for different marker proteins, so that all marker proteins are captured in the same column.

A sample from a subject is typically brought into contact with the surface under conditions suitable for binding of marker proteins in the sample to the surface. The proteins present in the sample may optionally be fractionated and the fraction(s) comprising the markers being detected may be collected and brought into contact with the surface. Unbound material is washed away using an appropriate solvent or buffer, such as phosphate buffered saline (PBS), designed to elute unbound proteins and other substances whilst retaining the markers of interest bound to the surface. The sample from the subject is typically a blood, plasma or serum sample.

The captured marker proteins may be detected by any suitable method. In one embodiment, bound markers may be detected by an immunoassay, for example by an ELISA assay or fluorescence-based immunoassay. In a typical immunoassay, the bound marker may be detected using an antibody, or fragment thereof, which will bind to the marker. Where the capture reagent is an antibody, the detector antibody is typically a different antibody to the capture reagent. Typically, the antibody binds the marker at a site which is different to the site which binds the capture reagent.

The antibody may be specific for the complex formed between the marker and the capture reagent immobilised on the support.

Generally, the antibody is labelled with a label that may be detected either directly or indirectly. A directly detectable label may comprise a fluorescent label such as fluorescein, Texas red, rhodamine or Oregon green. The binding of a fluorescently labelled antibody to the immobilised capture reagent/marker complex may be detected by microscopy. For example, using a fluorescent, bifocal or confocal microscope.
Preferably, the antibody is conjugated to a label that may be detected indirectly. The label that may be detected indirectly may comprise an enzyme which acts on a precipitating non-fluorescent substrate that can be detected using an automated reader. An automated reader is typically based on a video camera and image analysis software. The automated reader is capable of providing a measure of the quantity of each detected marker. Preferred enzymes include alkaline phosphatase and horseradish peroxidase. Automated readers are well known in the art and include, for example the Grifols Tritorus analyser (Grifols, Cambridge UK).

Other indirect methods may be used to enhance the signal from the detector antibody. For example, the detector antibody may be biotinylated allowing detection using streptavidin conjugated to an enzyme such as alkaline phosphatase or horseradish peroxidase or streptavidin conjugated to a fluorescent probe such as FITC or Texas red.

In all detection steps, it is desirable to include an agent to minimise non-specific binding of the second and subsequent agent. For example bovine serum albumin (BSA) or foetal calf serum (FCS) may be used to block non-specific binding.

In one embodiment, the captured proteins may be detected by gas phase ion spectrometry, such as mass spectrometry, for example MALDI or SELDI, following elution of the proteins from the surface, e.g. chip or beads. Such detection methods enable different proteins and different forms of the same protein to be distinguished without the need for labelling.

Gas phase ion spectrometry requires a gas phase ion spectrometer to detect gas phase ions. Gas phase ion spectrometers include an ion source that supplies gas phase ions and include mass spectrometers, ion mobility spectrometers and total ion current measuring devices. A mass spectrometer is a gas phase ion spectrometer that measures a parameter which can be translated into mass-to-charge rations of gas phase ions. Mass spectrometers typically include an ion source and a mass analyser. Examples of mass spectrometers are time-of-flight (ToF), magnetic sector, quadrupole filter, ion trap, ion cyclotron resonance, electrostatic sector analyser and hybrids of these. A laser desorption mass spectrometer is a mass spectrometer which uses laser as a means to desorb, volatilize and ionize an analyte. A tandem mass
spectrometer is mass spectrometer that is capable of performing two successive stages of m/z-based discrimination or measurement of ions, including ions in an ion mixture.

The captured markers may be desorbed or ionized from the capture surface using any suitable source of ionizing energy, such as high energy particles generated via beta decay of radionuclides or primary ions generating secondary ions. The preferred form of ionizing energy for solid phase analytes is a laser.

A preferred mass spectrometric technique for use in the invention is SELDI (Surface Enhanced Laser Desorption and Ionization) which is a method of desorption/ionization gas phase ion spectrometry in which the marker proteins are captured on the surface of a protein chip, or SELDI probe, that engages the probe interface of the gas phase ion spectrometer. In this embodiment using a protein chip array to capture the marker proteins, a protein chip reader may be used to detect the bound markers. Proteins bound on the protein chip are typically allowed to dry prior to the addition of an energy absorbing molecule (EAM) solution and the insertion of the protein chip into a protein chip reader to measure the molecular weights of the bound proteins. Upon laser activation in the protein chip reader, the sample becomes irradiated and the adsorption/ionization proceeds to liberate gaseous ions from the protein chip arrays. These gaseous ions enter the time of flight mass spectrometry (ToF MS) region of the protein chip reader which measures the mass-to-charge ratio (m/z) of each protein, based on its velocity through an ion chamber. Time lag focusing may be used to increase the mass accuracy of the signal output. Signal processing is accomplished by high speed analogue to digital converter, which is linked to a personal computer. Detected proteins are displayed as a series of peaks. The amplitude of the peaks is an indication of the amount of each protein present in a sample. Suitable EAMs for use in methods of the invention include cinnamic acid derivatives, sinapinic acid and dihydroxybenzoic acid.

Expression data may also be obtained by nephelometry. Nephelometry is a laboratory technique used to obtain a measurement of the amount of a marker accurately and rapidly. The data may, for example, be obtained by particle-enhanced immunonephelometry or rate nephelometry. The BNII analyser (Dade Behring, Milton Keynes, UK) is suitable for performing particle enhanced
immunonephelemy. The Beckman Immage (Beckman Coniter, High Wycombe, UK) may be used to perform rate nephelemy. The Beckman Immage may be calibrated against the International Reference Preparation CRM 470. Measurement of marker expression may be carried out by following the instructions provided by the manufacturer of the analyser used.

Other detection methods that may be used include optical techniques, such as confocal or fluorescence microscopy, electrochemical techniques, such as voltametry and amperometry, atomic force microscopy and radio frequency techniques, such as multipolar resonance spectroscopy.

The expression pattern of the markers of interest is examined to determine whether expression of the markers is indicative of the patient having TB. Any suitable method of analysis may be used. Typically, the analysis method used comprises comparing the expression data obtained from a subject to expression data obtained from patients known to have TB and control subjects who do not have a *Mycobacterium tuberculosis* infection. It can then be determined whether or not the expression of the markers in the subject is more similar to the expression pattern observed in known TB patients or to the expression pattern observed in control subjects. The method of analysis typically measures the likelihood of a subject having TB.

The patients having TB have typically been diagnosed as having TB as a result of culture of *Mycobacterium tuberculosis* from a sample derived from each patient. The control subjects may be selected from one or more of patients with respiratory infections other than TB, patients with sarcoidosis, patients with inflammatory bowel disease, patients with malaria, patients with human African trypanosomiasis (HAT), patients with neurological diseases, patients with autoimmune disease, patients with melanoma and healthy subjects. Patients suffering from other diseases not listed above, which patients do not have TB may also be used as control subjects. Typically, the control subject expression data to which the expression pattern of markers from the test subjects are compared comprise at least two, for example at least three, at least four, at least five, at least six, at least seven or at least eight, of the above mentioned subjects. Patients who are
HIV positive are particularly susceptible to disease. The TB patients and/or the control subjects may be HIV positive or HIV negative.

The TB and control samples may be taken from patients and/or subjects from more than one, for example, two or more, three or more, four or more, five or more, eight or more or ten or more, geographical sites. Each geographical site may be a different continent, country or region within a country. Different samples from TB and/or control subjects may be processed to obtain expression data at different times. For example, the samples may be obtained and/or processed over any suitable period of time, such as one month to two years, three months to eighteen months or six months to one year.

The method by which it is determined whether the expression data is indicative of TB, or not, is typically implemented using a computer. The computer may be physically separate from or may be coupled to the reader used to generate expression data, for example to the mass spectrometer.

Supervised machine learning classification methods may be used to discriminate the expression data of patients with TB from expression data of the control subjects. The machine learning classifier is first trained using training expression data from TB patients and training control data from the control subjects.

A method of training a machine learning classifier to distinguish expression data from a TB patient from expression data from a subject who does not have TB is illustrated in the flow chart of Figure 1. The steps carried out by a computer program executed on a computer system are illustrated schematically by a dotted line in Figure 1. The training data from TB patients and control objects (data D1) represent input variables (typically m/z values, ELISA values or nephelometry values). In step S1, the computer maps these input variables to feature space using a kernel and in step S2 the classifier learns to discriminate between TB data and control data thus producing a training classifier, such as a SVM, to discriminate between TB data and control data.

The trained classifier may then be tested using expression data from further TB patients and further control subjects. A method of testing the generalisation of a machine learning classifier is illustrated in the flow chart of Figure 2. The computer-implemented steps are illustrated schematically by a dotted line in Figure 2.
Independent training and testing sets may be used, with similar numbers of TB cases and controls and similar representation of age and sex in each set, for example as shown in Table 1. The testing data from TB patients and/or control subjects (data D2) represent input variables (typically \( m/z \) values, ELISA values or nephelometry values). The computer maps these input variables to feature space using a kernel in step S3 and the classifier produced using training data is used in step S4 to assign the class of the input variables as being TB data or non-TB data. It can then be determined whether the test data has been classified correctly or mis-classified.

A trained machine learning classifier may be used to determine whether expression data from a subject whom it is wished to diagnose as having, or not having, TB is indicative of the patient having, or not having, TB. The trained machine learning classifier used in such a method of diagnosis may have been tested as described above, but this testing step is not essential. Figure 3 is a flow chart which illustrates a computer-implemented method of diagnosis according to the invention. The computer-implemented steps are illustrated schematically in Figure 3 by a dotted line. The data from the test subject (i.e. a new unknown subject) labelled D3 in Figure 3 represents the input variables. In step S5, the computer maps the input variables (typically \( m/z \) values, ELISA values or nephelometry values) to feature space using a kernel and the previously obtained classifier is used in step S6 to classify the sample as being a TB sample or non-TB sample. Hence, the test subject is diagnosed as having or not having TB.

Suitable machine learning classifiers include the single layer perceptron (SLP), the multi-layer perceptron (MLP), decision trees and support vector machines. Preferably the classifier in a support vector machine. More preferably, the classifier is a Gaussian kernel support vector machine.

A supervised learning algorithm is tasked to find a decision function capable of assigning the correct label for a set of input/output pairs of examples, called the training data. The ability of the decision function to predict correct labels for unseen samples (test data) is known as its generalization. Current machine learning methods such as support vector machines (SVM) aim to optimize this property. The generalization of a classifier is dependent on a set of parameters (model) that must be chosen to optimise performance. For this purpose a grid search strategy may be
adopted in which a range of parameter values are discretized and tested using cross-validation.

A dataset $D$ is represented by a sample of input vectors, $X$, (i.e. exemplars of categories) with their corresponding sample of output labels, $Y$, $D = [X,Y]$. A sample input vector is represented by $x$. The mass spectrum of the $i$-th sample is represented as an $n$-dimensional (number of mass clusters) vector $x_i$ with an associated class label $y_i$ ($+1$ for TB, $-1$ for control) where $i = 1,..,m$ and $m$ is the number of samples. The spectrum vector elements are denoted by $x_{ik}$ where $i = 1,..,m$ and $k = 1,..,n$. The classifier prediction of a sample class label $y_i$ is denoted by $\hat{y}_i$.

The Support Vector Machine (SVM) maps its inputs to a high or even infinite dimensional feature space. The output of the SVM is then a linear thresholded function of the mapped inputs in the feature space, which may be nonlinear in the original input space. The mapping is accomplished by a user-selected reproducing kernel function $K(x,x')$ where $x$ and $x'$ are input vectors. The kernel function must satisfy Mercer’s conditions. Well-known examples of kernels include the Gaussian

$$K(x,x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$$

where the parameter $\sigma$ determines the width; and the polynomial

$$K(x,x') = (x \cdot x')^d$$

where $d$ determines the degree. When $d = 1$ it is called the linear kernel and corresponds to the identity map of the input data. A trained SVM classifier has the form $\text{svm\_classifier}(x) = \text{sign} \left( \sum_{i=1}^{m} \alpha_i K(x_i, x) + b \right)$ and training determines the values of $\alpha$ and $b$. Typically, many of the $\alpha$s will be zero. Those that are non-zero are called 'support vectors' and are used to define a separation hyperplane in the transformed feature space. Training a SVM is a convex (quadratic) optimization problem not subject to local minima unlike a multi-layer perceptron. There are many packages available to train an SVM; such as $SVM^\text{light}$ (Joachims, 1999) and, in particular, soft-margin SVMs which are practicable when data are noisy. In this case the algorithm also minimizes the distance of incorrectly classified examples to the margin by adjusting a penalty value, $C$, called the soft-margin parameter.
The Single Layer Perceptron (SLP) (Rosenblatt, 1962) is an artificial neural network with one output neuron that computes a linear combination of the values given by the input layer. The discrimination function is given by

\[ \hat{y} = \text{sign} \left( \sum_{i=1}^{n} w_i x_i + b \right) \]

where weights \( w \) are obtained by an iterative learning algorithm designed to reduce the total classification error \( \sum_{i=1}^{m} |y_i - \hat{y}_i| \).

The Multi-Layer Perceptron (MLP) (McClelland and Rumelhart, 1986) is a generalization of the SLP with intermediate layers of hidden neurons. It tackles the problem of non-linearly separable classes by allowing the neurons to process their inputs with a sigmoid function on the activation level \( f(a) = \frac{1}{1 + e^{-a}} \). In this network the weights are learned by a back-propagation algorithm which is a gradient descent rule to minimize the error given by \( \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \).

A decision tree learns to classify a dataset of samples \( D = \{X, Y\} \) by aggregating their features within a set of nodes organized in a binary tree structure. To find the tree structure, sample features are tested according to their discriminative power using a splitting criterion: for a given mass peak \( x_{ik} \) the test \( x_{ik} < T \) where \( T \) is any test that produces a binary partition of dataset \( D \). In the C4.5 (Quinlan et al., 1993) classifier the test thresholds are evaluated by an information-gain splitting criterion

\[ \text{Gain}(D, T) = \text{Info}(D) - \sum_{i=1}^{z} \left( \frac{|D_i|}{|D|} \times \text{Info}(D_i) \right) \]

where \( \text{Info}(D) \) is an entropy measure of the class to which the sample belongs and \( z \) is the number of outcomes of the test \( T \). An iterative algorithm places nodes with increasing information gain from the root to the leaves of the tree. The final tree might be pruned in order to get a more compact representation of the classifier. A testing set sample can be classified by testing its mass peak values against those in the nodes of the tree following a path from the root to a leaf with a classification output. The C5.0 algorithm is an extended version of C4.5 that winnows irrelevant features and incorporates variable misclassification costs (http://www.rulequest.com/). The Alternating Decision Tree (ADTree) (Freund
and Mason, 1999) is a tree with additional nodes for predicting values that are summed over a classification path and the final output is the sign of this sum.

Any suitable cross-validation scheme may be used such as $k$-fold cross-validation or $k$-fold cross-validation with test. In $k$-fold cross-validation the training set is randomly split in $k$ groups of equally distributed positive and negative cases. A classifier is trained on $k-1$ of the groups and its generalization performance is validated on the remaining group. This process is repeated $k$ times, each time holding out a different validation subset and the average represents the overall generalization. In the second scheme, $k$-fold cross-validation with test, the data is first randomly split into training and testing sets. A $k$-fold cross-validation is performed on the training set and the generalization is obtained on the unseen testing set.

The generalization performance of the classifiers may be assessed by considering the number of correctly classified (true positives, TP and true negatives, TN) and incorrectly classified (false positives, FP and false negatives, FN) cases in the testing set. Sensitivity ($se$), may be defined as the conditional probability of a true positive $se = TP/(TP + FN)$, specificity ($sp$) as the conditional probability of a true negative $sp = TN/(TN + FP)$, and accuracy ($ac$) as the proportion of correct classifications $ac = (TP + TN)/(TP + FP + TN + FN)$. The performance of a classifier expressed by its true positive rate ($se$) and false positive rate ($1 - sp$) can be plotted in a receiver operator curve (ROC) space.

Robust estimates of the generalization capability of the classifier may be provided by carrying out 10-fold cross-validation with test. For example, one hundred 80:20 train:test sets may be generated by random sampling without replacement in the entire dataset. For each 80:20 train:test set a 10-fold cross validation is carried out on the training set and the parameter with the best performance is chosen. The SVM may be re-trained with the best parameter over all the 10 subsets and the final performance is assessed on the testing set. Each ROC curve may be smoothed, sampled and averaged in order to show the mean curve with standard deviation.

The invention further provides a computer-implemented method of diagnosing TB, said method consisting essentially of the steps of:

(a) inputting expression data of two or more markers in a subject; and
(b) determining whether expression of said markers is indicative of TB using a computer system programmed with a trained support vector machine (SVM);

thereby diagnosing whether or not said patient has TB.

The expression data may relate to any two or more markers which are differentially expressed in TB patients and control subjects and include the markers described above. In one embodiment, the expression data is a proteomic profile from a sample from the subject, typically a blood, plasma or serum sample, obtained by SELDI analysis.

The support vector machine is trained as described above and is preferably a Gaussian kernel support vector machine. The computer system programmed with the trained support vector machine classifies the expression data from the subject as being indicative of the subject having TB, or of the subject not having TB. Accordingly, the output from the computer system enables diagnosis of the subject as having, or not having, TB.

Based on a diagnosis of TB by a method of the invention, further processes may be instigated. A method of diagnosis according to the invention may further comprise administering to a patient diagnosed as having TB, a medicament for the treatment of TB. A medicament for treating TB is a substance or composition that, when administered to a subject in a therapeutically effective amount, alleviates the symptoms or otherwise lessens the suffering of the subject. The substance or composition may be an agent which kills or disables Mycobacterium tuberculosis, for example by preventing its replication. Suitable medicaments include isoniazid, rifampin, pyrazinamide and ethambutol. The exact treatment regime may depend on the state of the individual, for example whether the individual is pregnant, HIV-seropositive, diabetic, etc and may readily be determined by a physician.

The present invention further provides a method of training a support vector machine (SVM) classifier to diagnose TB, said method consisting essentially of the steps of:

(a) providing training data which comprises:

(i) training data relating to two or more markers in each of a first set of TB patients; and
(ii) training data relating to said two or more markers in each of a first set of control subjects; and

(b) using a SVM to discriminate the training data of TB patients from the training data of control subjects;

thereby training the SVM to diagnose TB.

The method optionally further consists essentially of:

(c) providing testing data which comprises:

(i) testing data relating to said two or more markers in each of a second set of TB patients; and

(ii) testing data relating to said two or more markers in each of a second set of control subjects;

(d) determining the ability of the SVM to correctly discriminate the testing data of TB patients from the testing data of control subjects.

The training and testing data may be obtained by any suitable method, such as those described above.

The testing data is typically used to determine the sensitivity, specificity and/or accuracy of the SVM classifier.

The invention further provides an apparatus arranged to perform a method of diagnosis according to the invention, which apparatus consists essentially of:

(i) means for receiving expression data of two or more markers in a sample from a subject;

(ii) a module for determining whether said data is indicative of TB, wherein said module comprises a trained machine learning classifier capable of distinguishing data from a TB patient and data from a control subject; and

(iii) means for indicating the results of said determination.

The means for receiving expression data may be a keyboard into which data may be entered manually. Alternatively, the expression data may be received directly from the computer analysing the expression data, such as the protein chip reader or automated image analyser. The expression data may be received by a wire, or by a wireless connection. As a further alternative, the expression data may be recorded on a storage medium in a form readable by the apparatus. The storage medium may be placed in a suitable reader comprised within the apparatus.
The training, testing and or expression data from a subject being tested for TB may be raw data or may be processed prior to being inputted into the computer system. The computer system may comprise a means for converting raw data into a form suitable for further analysis.

The module for determining whether the data is indicative of TB, comprises a machine learning classifier which has been trained by a method as described herein such that it is able to distinguish expression data characteristic of a TB patient from expression data characteristic of a control subject.

The means for indicating the results of said determination may be a visual screen, audio output or printout. The results typically indicate the classification of the expression data and may optionally indicate a degree of certainty that the classification is correct.

The apparatus of the invention may be a personal computer. The personal computer may be a laptop. Alternatively, the apparatus may be a hand held computer, for example a specifically designed hand held computer, which has the advantage of being readily transportable in the field.

The invention further provides a computer program executable by a computer system, the computer program being capable, on execution by the computer system, of causing the computer system to perform a method of diagnosis according to the invention. The computer program generally comprises a machine learning classifier, preferably a support vector machine, which has been trained as described herein.

The invention further provides a storage medium storing in a form readable by a computer system a computer program of the invention. Any suitable storage medium may be used such as a CD-ROM or floppy disk.

In a further aspect, the invention provides a kit for use in the diagnosis of TB. The kit typically comprises means for detecting two or more markers as defined herein. The means of detection typically comprises a capture surface as described herein, such as a protein chip or array of specific binding reagents such as antibodies or antibody fragments. The kit may comprise instructions for operation in the form of a label or separate insert. For example, the instructions may inform a consumer how to collect the sample, how to incubate the sample with the capture surface and/or how to wash the probe. The kit may comprise instructions for inputting
expression data of the markers into an apparatus of the invention. The kit may
comprise a storage medium of the invention.

The kit is preferably adapted to detect any combination of two or more, such
as three, four, five or six or more of the markers, transthyretin, neopterin, CRP, SAA,
Apo-A1, serum albumin, Apo-A2, hemoglobin beta, haptoglobin protein, DEP
domain protein, A2GL and hypothetical protein DFKZp6671032. In one preferred
embodiment, the kit is adapted to detect any combination of two or more, such as
three or four of the markers transthyretin, neopterin, CRP and SAA, for example,
transthyretin, neopterin and CRP. The kit may be capable of detecting additional
markers other than these four specified markers.

The kit may be adapted to detect the positive markers and/or negative
markers set out in the Table below.

<table>
<thead>
<tr>
<th>Positively Correlated</th>
<th>Negatively Correlated</th>
</tr>
</thead>
<tbody>
<tr>
<td>'M18394_9'</td>
<td>'M4100_03'</td>
</tr>
<tr>
<td>'M8952_75'</td>
<td>'M3898_52'</td>
</tr>
<tr>
<td>'M11720_0'</td>
<td>'M13774_3'</td>
</tr>
<tr>
<td>'M11454_1'</td>
<td>'M13972_1'</td>
</tr>
<tr>
<td>'M18591_2'</td>
<td>'M3322_01'</td>
</tr>
<tr>
<td>'M11488_1'</td>
<td>'M2956_45'</td>
</tr>
<tr>
<td>'M11541_5'</td>
<td>'M5644_96'</td>
</tr>
<tr>
<td>'M9076_68'</td>
<td>'M3939_63'</td>
</tr>
<tr>
<td>'M8895_13'</td>
<td>'M4056_39'</td>
</tr>
<tr>
<td>'M10856_8'</td>
<td>'M6649_74'</td>
</tr>
</tbody>
</table>

In this embodiment, the detection means is preferably a protein chip.

The kit may additionally comprise one or more sample of one or more marker
in a container. The marker provided in the kit may be used as a control or for
calibration.

The invention also provides methods for identifying candidate agents for the
treatment of TB. Candidate agents may be identified by assaying for activity of a
test agent in modifying activity or expression of one or more of transthyretin,
neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2, hemoglobin beta,
haptoglobin, DEP domain protein or A2GL. The biological activities of each of
transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2, hemoglobin beta, haptoglobin, DEP domain protein or A2GL are known in the art. Accordingly, the skilled person would readily be able to perform assays to assess the effect of a test agent on the activity of any one of transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2, hemoglobin beta, haptoglobin, DEP domain protein or A2GL.

In one embodiment of the invention, candidate therapeutic agents may be identified by determining the effect of a test agent on the expression of one or more TB marker in cells infected with *Mycobacterium tuberculosis*. The one or more TB marker is generally selected from transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2, hemoglobin beta, haptoglobin, DEP domain protein, A2GL and hypothetical protein DFKZp667I032. An increase or decrease in expression of one or more marker indicates that the test agent is useful in the treatment of TB. Typically, where the marker is a positive marker of TB, a test agent useful in treating TB reduces the level of expression of the marker compared to the level of expression in infected cells in the absence of the test agent. Typically, where the marker is a negative marker of TB, a test agent useful in treating TB increases the level of expression of the marker compared to the level of expression in infected cells in the absence of the test agent.

The infected cells may be *in vivo* or *ex vivo*. Where the cells are *in vivo*, they are typically present in an experimental animal, typically a rodent, such as a mouse or a rat. The infected cells may be any cells which *Mycobacterium tuberculosis* is capable of infecting. In one embodiment the cells are cells of the respiratory system, or cell lines derived therefrom.

Also provided by the invention are candidate therapeutic agents identified by such methods of the invention. Suitable candidate agents include antibodies specific for one of transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2, hemoglobin beta, haptoglobin, DEP domain protein or A2GL.

The following Examples illustrate the invention.
Examples

Example 1: Selection of patients and control subjects

To develop new approaches for diagnosing TB we collected sera from cases (n=179) and controls (n=170) from multiple sites (UK, Angola, The Gambia and Uganda) representing patients from at least 4 ethnic backgrounds (Table 1). We confined ourselves to patients with TB who presented with typical manifestations of pulmonary disease (Rathman et al., 2003), because this is the commonest presentation of adult TB in all geographic areas. Diagnosis was confirmed by culture of M. tuberculosis. Details of patients that include both smear positive and smear negative cases, and control subjects (including HIV status) are given in Tables 1 and 2a. As expected, most patients presented with cough, fever and weight loss, and the majority had cavitary pulmonary disease.

For our control subjects, we recruited healthy volunteers as well as patients having conditions with clinical features that can overlap with TB (Table 2b). Our control subjects have heterogeneous causes of inflammation that have been confirmed by standard diagnostic criteria. For example, we included patients with sarcoidosis, which is frequently included in the differential diagnosis of pulmonary TB, and other severe respiratory infections representing patients who have non-tuberculous destructive pulmonary pathology. To allow for systemic inflammatory processes that can mimic TB, we recruited patients with other systemic infections as well as patients with inflammatory bowel disease and autoimmune diseases.

Example 2: Proteomic profiling and supervised machine learning classification

We first profiled 349 serum samples from these subjects on weak exchange (CM10) protein chip arrays by Surface Enhanced Laser Desorption Ionisation Time of Flight Mass Spectrometry (SELDI-ToF MS) (Issaq et al., 2002; von Eggeling et al., 2001) and identified 219 peak clusters from m/z spectra in the range 2,000-100,000. We then used state-of-the-art supervised machine learning classification methods (Table 3 and Figure 4) to discriminate the proteomic spectra of patients with TB from the controls using the training-testing-set approach (Table 1). The ability of a classifier to correctly discriminate data in the testing set is known as its generalization performance (Vapnik, 1998; Cristianini and Shawe-Taylor,
2000). We compared the generalization performance of a variety of classifiers by plotting their performance on such a testing set in Receiver Operating Characteristic (ROC) space.

In our study the SLP did not provide an optimal discriminative function, giving an accuracy of 86.5% in the independent test set (Table 3). With our data the MLP showed similar generalization performance to SLP, classifying with an accuracy of 86.5% (Table 3). In the TB versus control dataset (Table 2) the ADTree and the C4.5 classifiers achieved accuracies of 92.3% and 91.0% respectively (Table 3), but relied on AdaBoost boosting to achieve such levels of generalization (Witten and Frank, 2000) (Table 3). We used AdaBoost with 100 iterations for the ADTree and C4.5 classifiers, and boosting with a maximum of 10 iterations for the non-commercial version of the C5.0 classifier.

A Gaussian kernel support vector machine (Boser et al., 1992; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000) (SVM, Table 3) is the best discriminator between TB and control groups, having a sensitivity of 93.5% and a specificity of 94.9% (overall accuracy 94.2%). Five TB samples and 4 controls in the testing set were misclassified. This SVM classifier defines the convex hull of the ROC space achieving the best accuracy.

We applied a further test of generalization performance of the SVM by carrying out 10-fold cross-validation on the entire set of spectra (both training and testing), obtaining accuracy of 93.1±3.8%, sensitivity of 94.4±4.5% and specificity of 91.8±8.8% when optimised for accuracy. We also evaluated the generalisation performance of the SVM by varying the proportions of train:test cases from 90:10 to 50:50. For 80:20 sets, we obtained values for accuracy, sensitivity and specificity exceeding 90%. The robustness of the SVM is further confirmed by its mean performance on 100 randomly generated 80:20 sets as shown in the ROC curve, with an area under the curve (AUC) of 0.96. Figure 5 shows the averaged ROC using the 10-fold train cross validation test. In Figure 5a the kernel parameter is selected on sensitivity only and in Figure 5b the kernel parameter is selected on specificity criteria.

In spite of the deliberate heterogeneity of the control group, our classifier discriminates accurately between patients with TB (both smear negative and smear
positive) and those with a range of infective and non-infective inflammatory conditions. These results show that TB is amenable to a proteomic-signature based diagnostic approach. Artefacts associated with sample collection, handling or spectrum generation could potentially create spurious classifications. However, interspersing the processing of samples from TB cases and control subjects over a 6 month period and using samples from 4 different geographic sites and varying HIV sero-status, makes systematic biases between cases and control subjects highly unlikely. As a measure of reproducibility of the mass spectra, 28 universal control spectra run at different times over a 6 month period were correctly classified as control subjects by the SVM classifier obtained in the 10-fold cross-validation. In a clinic population where the prevalence of TB in patients presenting with respiratory symptoms is around 10%, the positive and negative predictive values for our best classifier would be 67% and 99% respectively. This diagnostic accuracy surpasses that of other available diagnostic options.

Example 3: Selection of markers

However, while SELDI technology can provide a diagnostic test for TB that makes no prior assumptions about the identities of proteins constituting an informative signature, cost and complexity may preclude its widespread general use. We therefore selected a subset of informative peak clusters for further evaluation by applying a correlation filter method to detect independently informative peaks (Guyon and Eliseeff, 2003). We ranked 10 mass clusters with the highest positive, and 10 with the highest negative, Pearson correlation coefficients. The m/z values of these markers is shown in the Table below.

<table>
<thead>
<tr>
<th>Positively Correlated</th>
<th>Negatively Correlated</th>
</tr>
</thead>
<tbody>
<tr>
<td>'M18394.9'</td>
<td>'M4100.03'</td>
</tr>
<tr>
<td>'M8952.75'</td>
<td>'M3898.52'</td>
</tr>
<tr>
<td>'M11720.0'</td>
<td>'M13774.3'</td>
</tr>
<tr>
<td>'M11454.1'</td>
<td>'M13972.1'</td>
</tr>
<tr>
<td>'M18591.2'</td>
<td>'M3322.01'</td>
</tr>
<tr>
<td>'M11488.1'</td>
<td>'M2956.45'</td>
</tr>
<tr>
<td>'M11541.5'</td>
<td>'M5644.96'</td>
</tr>
<tr>
<td>'M9076.68'</td>
<td>'M3939.63'</td>
</tr>
<tr>
<td>'M8895.13'</td>
<td>'M4056.39'</td>
</tr>
<tr>
<td>'M10856.8'</td>
<td>'M6649.74'</td>
</tr>
</tbody>
</table>
To study the discriminatory power of the selected 20 mass clusters we first paired each mass with every other (400 pairs) and trained SVM classifiers to diagnose TB cases. The results are shown in Table 4. We ranked generalization performance by accuracy and showed that 20 pairs (5%) of selected mass clusters gave accuracies greater than 80% and 17 of these combined negatively-correlated and positively-correlated mass clusters. No mass cluster pair achieved sensitivities and specificities greater than 95% and 85%, respectively, confirming that better generalization relies on combinations of more than two mass peaks. Second, an SVM trained with just the 20 correlation-selected mass clusters achieved an accuracy of 89.7% on the independent test set indicating that these clusters contain most relevant discriminatory information. Information in remaining peak clusters (n = 199) retains an inferior though acceptable diagnostic accuracy (85.9%). We summarised the generalization performance of the SVMs in ROC space using different sets of mass clusters. The ROC convex hull is defined by 2 classifiers. The highest specificity was obtained with all peaks minus the 10 that were positively correlated (i.e. 209 in total), confirming information value in negatively correlated peaks. The other optimal classifier was obtained after using only 10 positively and 10 negatively correlated subsets of mass clusters.

Example 4: Identification of markers

Using high-resolution mass-spectrometry after tryptic digestion we identified an 11.5kDa ‘positive’ marker and a 13.7kDa ‘negative’ marker as the des-arginine variant of serum amyloid A1 (SAA1) and transthyretin, respectively. Interestingly, these peptides, selected by Pearson correlation analysis and confirmed by SVM classification of proteomic signatures, have already been independently associated with pathophysiological processes in TB. SAA is an acute phase protein associated with circulating high-density lipoprotein (HDL) (Kierman et al., 2003) and modulating lipid trafficking and immune responses. It is the precursor protein in reactive amyloidosis, which complicates chronic TB in some individuals, and is a marker of disease activity in several inflammatory states including tuberculosis (Salazar et al., 2001). Transthyretin is a 55kDa homotetramer in serum and a major transporter of thyroxine and tri-iodothyronine, as well as vitamin A (retinol or trans-
retinoic acid) through association with retinol-binding protein (Peterson, 1971). Retinoic acid stimulates monocyte differentiation and inhibits multiplication of *M. tuberculosis* in human macrophages (Crowle *et al.*, 1989). Low levels of vitamin A, correlating with reduced transthyretin and elevated C-reactive protein levels, have been reported in patients with TB (Hanekom *et al.*, 1997; Koyanagi *et al.*, 2004).

**Example 5: Immunoassay tests and supervised machine learning classification**

To translate from proteomic signatures to conventional test formats, we quantitated serum SAA and transthyretin by immunoassay in all subjects. Because both peptides are markers of inflammation, we also measured C-reactive protein (CRP) and neopterin that have previously been used to monitor disease activity in TB (Hosp *et al.*, 1997). We then parameterised polynomial and Gaussian kernel SVMs for these 4 markers. The best 4 classifiers were obtained using Gaussian SVMs. The SVM classifier trained with transthyretin, CRP and neopterin values discriminated TB from control patients with an accuracy of 84% (82% sensitivity, 86% specificity). Other optimised classifiers were with SAA and CRP with transthyretin included, and using transthyretin and neopterin. Inclusion of additional markers in the original signature is likely to improve accuracy of immunoassay-based classifications.

A truncated form of transthyretin is a negative marker in proteomic fingerprinting studies on ovarian cancer (Zhang *et al.*, 2004) and SAA is a positive marker in Severe Acute Respiratory Syndrome (SARS) (Ren *et al.*, 2004) and indicates relapse in nasopharyngeal cancer (Cho *et al.*, 2004). Although single protein markers may have insufficient accuracy in the diagnosis of TB, the use of proteome-guided analysis coupled with machine learning methods such as SVM can achieve accuracies that are superior to current standard methods. These findings suggest that markers with low individual diagnostic specificities can boost diagnostic yields when used in particular combinations. In some cases, truncated or fragmented derivatives of common plasma proteins may be more specific markers of some diseases and arise by proteolytic enzyme induction characteristic of defined disease states (Tolson *et al.*, 2004). Preservation of high diagnostic accuracy when translating from proteomic signatures to immunoassays, and the biological plausibility of identified biomarkers establishes the value of SVM classifiers for
diagnosis of TB and provides strong foundations for serological testing. Provision of trained SVM classifiers on personal computers provides an opportunity to aid TB diagnosis using immunoassays (or where available, SELDI proteomic analysis). These tests can then be applied to longitudinal studies of TB and other difficult diagnostic categories such as patients with sputum negative TB, extra-pulmonary cases and paediatric infections.

Example 6: Materials and methods

Serum collection and storage. Serum samples (179) were collected from patients with retrospectively confirmed culture-positive TB (Table 2). Banked sera collected in Uganda and The Gambia were obtained from the World Health Organisation TB specimen bank (http://www.who.int/tdr/diseases/tb/specimen.htm), and others were collected prospectively from patients presenting with TB to the inpatient and outpatient facilities at St George’s Hospital, London, UK. Serum samples (170) from control patients with a range of other inflammatory conditions were collected at St George’s Hospital, UK, the Angotrip treatment centre, Angola and The Gambia. Fully informed consent was obtained in each case, in accordance with local Research Ethical Committee policy. Clinical information was archived in a linked, anonymised database. Serum was separated from 5ml blood by centrifugation, and samples allowed to clot for 30 minutes at room temperature in sterile glass tubes. Aliquots (100μl) were frozen (-80°C) within 1 hour of collection, and subjected to no more than two freeze-thaw cycles prior to mass spectrum analysis.

Sample preparation for mass spectrometry. Samples were applied to CM10 protein chip arrays (Ciphergen, Fremont, CA, USA) as described previously (Papadopoulos et al., 2004), and a saturated solution of sinapinic acid in 50% acetonitrile, 0.5% trifluoroacetic acid was applied twice to each spot on the array, with air drying between each application. To minimise bias, sera from TB patients and controls were assayed on the same chips.

Surface Enhanced laser Desorption Ionisation Time of Flight Mass Spectrometry (SELDI-ToF MS). Time-of-flight spectra were generated using a PBS-II Mass spectrometer (Ciphergen, Freemont, CA, USA) at laser intensities of
200, 220 and 240, high mass 100kDa, detector sensitivity 8 and focus mass 10kDa. Each spot on the array was analysed from position 20 to 80, delta 4, with 7 shots per position, preceded by 2 warming shots at laser intensities of 205, 225 or 245. Each protein chip array included a ‘universal control’ sample (aliquoted from a single collection from one individual and stored at -80°C). Both groups of spectra (TB and controls) comprised samples run on different occasions over a 6 month period.

**Peak identification.** Spectra were calibrated weekly using the Ciphergen all-in-one protein and peptide calibrants, and normalised to the total ion current in the m/z range over 2,000-100,000 after baseline subtraction. For each patient a single spectrum generated at a laser intensity of 200, 220 or 240 was selected to minimise deviation of the total ion current to within 0.4-2.6 times the mean of all patients as described previously (Papadopoulos et al., 2004). Biomarker Wizard version 3.1 was used to identify corresponding peaks in each spectrum (‘peak clusters’) within 0.6% of the molecular mass. Signal-to-noise ratio was set at 10 for the first pass and 2 for the second pass. To assess reproducibility, coefficients of variation for peak size for spectra derived from a single sample run 25 times (6 assays) were 15.6% (intra-assay) and 24.4% (inter-assay). These data were obtained by averaging values for 9 of the highest amplitude peaks at the following m/z values: 5648, 6203, 6449, 6647, 8907, 9213, 9310, 9370 and 9419.

**Protein identification.** Serum (20µl) was incubated on ice (20 minutes) with 30µl denaturation buffer, diluted in 50µl binding buffer (denaturation buffer diluted 1:9 in 50mM Tris-HCl pH9.0) followed by a further 30 minute incubation on ice. Samples were applied to Q Ceramic HyperD spin columns (Ciphergen, 20 minutes), pre-equilibrated first in Tris (50mM, pH 9), followed by binding buffer. Both the 11.5kDa and 13.7kDa biomarkers were eluted from the spin column in elution buffer (50mM Na citrate, 0.1% octyl glucopyranoside, pH 3) and selective enrichment was confirmed by SELDI-ToF MS analysis of a sample of eluate applied to a CM10 protein chip array under conditions as described above for unfractionated serum.

The biomarkers were isolated by 1D SDS-PAGE (NuPAGE, 4-12% Bis-Tris, Invitrogen), stained with Coomassie Blue and excised from the gel. The gel pieces were washed three times in a mixture of ammonium bicarbonate (50mM) and acetonitrile (50%), dehydrated in acetonitrile (100%) and dried.
Proteins were subjected to in-gel trypsic digestion (15 minutes, RT) by the addition of trypsin (20ng/μl) in acetonitrile (10%) and ammonium bicarbonate (25mM), followed by a final incubation in ammonium bicarbonate (25mM) for 4 hours.

Peptide mass fingerprints (PMFs) of the digests were analysed by MALDI-ToF MS using 20% α-cyano-4-hydroxy-cinnamic acid (CHCA) as matrix. The results of the in-gel trypsic digest were corroborated by trypsic digestion following passive elution of the protein from the gel.

The PMFs were used to interrogate the MASCOT database which identified the peptides as having been derived, in one case from serum amyloid A1 (SAA1) and in the other, from transthyretin. The molecular weight observed in the mass spectrum (13.7kDa) for the protein identified as transthyretin corresponded closely to the theoretical value (13.76kDa) of this protein. However that observed for SAA1 (11.52kDa) was 156Da lower than its theoretical value (11.68kDa) suggesting that the protein was a SAA1 variant.

In order to investigate the nature of this variant, the trypsic digest was analysed in more detail and found to include a peptide at m/z 1551 that did not correspond to a trypsic peptide predicted from the full amino acid sequence of SAA1. It did, however, correspond to the 2-15 peptide (SFFSFLGEAFDGFAR) which would have resulted from loss of the N-terminal arginine.

**Immuniquantitation of biomarkers.** The lower limit detection for each marker and the antibody type used for detection were as follows: 0.7mg/l SAA with particle enhanced sheep anti-SAA, 1mg/l CRP with goat anti-CRP, 0.05g/l transthyretin with goat anti-transthyretin and 1.5nmol/l neopterin with rabbit anti-neopterin. Neopterin was measured by competitive ELISA using a kit (ELItest Neopterin, B.R.A.H.M.S Aktiengesellschaft, Germany) in a Triturus analyser (Grifols UK Ltd). Rate nephelometry was used for measurement of C-reactive protein, transthyretin (Beckmann Immage 800 analyser, Beckman Coulter UK, Ltd) and serum amyloid A (N latex SAA, BN II analyser, Dade-Behring, Marburg, Germany). The antibody used in the SAA assay detects total SAA. Values from ELISAs were scaled in the range 0-1 before use in SVM classification experiments, and all possible combinations were used as feature space.
Supervised Machine Learning. A dataset $D$ is represented by a sample of input vectors, $X$, (i.e. exemplars of categories) with their corresponding sample of output labels, $Y$, $D = [X,Y]$. A sample input vector is represented by $x$. The mass spectrum of the $i$-th sample is represented as an n-dimensional (number of mass clusters) vector $x_i$ with an associated class label $y_i$ (+1 for TB, -1 for control) where $i=1,...,m$ and $m$ is the number of samples. The spectrum vector elements are denoted by $x_{i,k}$ where $i=1,...,m$ and $k=1,...,n$. The classifier prediction of a sample class label $y_i$ is denoted by $\hat{y}_i$.

A supervised learning algorithm is tasked to find a decision function capable of assigning the correct label for a set of input/output pairs of examples, called the training data. The ability of the decision function to predict correct labels for unseen samples (test data) is know as its generalization. Current machine learning methods such as SVM aim to optimize this property. The generalization of a classifier is dependent on a set of parameters (model) that must be chosen to optimise performance. For this purpose we adopted a grid search strategy in which a range of parameters values are discretized and tested using cross-validation.

The Support Vector Machine (SVM) maps its inputs to a high or even infinite dimensional feature space (Vapnik et al., 1998; Aronszajn, 1950). The output of the SVM is then a linear thresholded function of the mapped inputs in the feature space, which may be nonlinear in the original input space. The mapping is accomplished by a user-selected reproducing kernel function $K(x,x')$ where $x$ and $x'$ are input vectors. The kernel function must satisfy Mercer’s conditions (Joachims, 1999). Well-known examples of kernels include the Gaussian $K(x,x') = e^{-\frac{|x-x'|^2}{2\sigma^2}}$ where the parameter $\sigma$ determines the width; and the polynomial $K(x,x') = (x \cdot x')^d$ where $d$ determines the degree. When $d = 1$ it is called the linear kernel and corresponds to the identity map of the input data. A trained SVM classifier has the form

$$\text{svm\_classifier}(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i K(x_i, x) + b\right)$$

and training determines the values of $\alpha$ and $b$. Typically, many of the $\alpha$'s will be zero. Those that are non-zero are called ‘support vectors’ and are used to define a separation hyperplane in the transformed
feature space. Training a SVM is a convex (quadratic) optimization problem not subject to local minima unlike a multi-layer perceptron. There are many packages available to train an SVM; we used SVM^light (Rosenblatt, 1962) and in particular we trained soft-margin SVMs which are practicable when data are noisy. In this case the algorithm also minimizes the distance of incorrectly classified examples to the margin by adjusting a penalty value, C, called the soft-margin parameter.

We used two cross-validation schemes. In k-fold cross-validation the training set is randomly split in k groups of equally distributed positive and negative cases. A classifier is trained on k–1 of the groups and its generalization performance is validated on the remaining group. This process is repeated k times, each time holding out a different validation subset and the average represents the overall generalization. In the second scheme, k-fold cross-validation with test, the data is first randomly split into training and testing sets. A k-fold cross-validation is performed on the training set and the generalization is obtained on the unseen testing set.

The generalization performance of the classifiers was assessed by considering the number of correctly classified (true positives, TP and true negatives, TN) and incorrectly classified (false positives, FP and false negatives, FN) cases in the testing set. Sensitivity (se), was defined as the conditional probability of a true positive se = TP/(TP + FN), specificity (sp) as the conditional probability of a true negative sp = TN/(TN + FP), and accuracy (ac) as the proportion of correct classifications ac = (TP + TN)/(TP + FP + TN + FN). The performance of a classifier expressed by its true positive rate (se) and false positive rate (1 – sp) can be plotted in a receiver operator curve (ROC) space.

We created independent training and testing sets, with similar numbers of TB cases and controls and similar representation of age and sex in each set (Table 1). Using these sets we evaluated the generalization performance of several supervised machine learning methods such as single layer perceptron (SLP) (McClelland and Rumelhart, 1986), multi layered perceptron (MLP) (Quinlan et al., 1993), tree classifiers (Freund and Mason, 1999; Freund and Schapire, 1996 and Witten and}

Frank, 2000) and support vector machines (Table 3).

To provide robust estimates of the generalization capability of the classifier we carried out 10-fold cross-validation with test. First, we generated one hundred
80:20 train:test sets by random sampling without replacement in the entire dataset. For each 80:20 train:test set a 10-fold c.v. is carried out on the training set and the parameter with the best performance is chosen. The SVM is re-trained with the best parameter over all the 10 subsets and the final performance is assessed on the testing set. In these experiments each ROC curve is smoothed, sampled and averaged in order to show the mean curve with standard deviation.

**Mass peak cluster selection.** We used the Pearson correlation coefficient to rank peaks for their discriminatory power. The Pearson correlation coefficient is defined as $R(k) = \frac{\text{covariance}(X_k, Y)}{\sqrt{\text{variance}(X_k)\text{variance}(Y)}}$ where $X_k$ is the random variable corresponding to the $k^{th}$ component of sample input vectors $x$ and $Y$ is the random variable of output labels.

The estimate of $R(k)$ is given by $\hat{R}(k) = \frac{\sum_{i=1}^{m} (x_{i,k} - \bar{x}_k)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_{i,k} - \bar{x}_k)^2 \sum_{i=1}^{m} (y_i - \bar{y})^2}}$, where $x_{i,k}$ correspond to value m/z of the mass cluster $k$ of sample $i$, $y_i$ is the class label for sample $i$ and $m$ is the number of samples. $R(i)$ may be used a test statistic to assess the significance of a variable and it is linked to the $t$-test. We calculated $\hat{R}(k)$ between values of each mass cluster and corresponding class labels across the training set (Table 1). We then used $\hat{R}(k)$ to rank positively and negatively correlated mass clusters. Using this approach we selected 10 mass clusters with the highest positive, and 10 with the highest negative, correlation coefficients. The decision boundary found by the classifier and discriminating mass cluster pairs in the feature space induced by the kernel is shown in Fig 2a (green lines).

**Software.** We used a chunking and decomposition implementation of the support vector machine SVM$^{light}$. We used Waikato Environment for Knowledge Analysis (WEKA) for decision tree algorithms, boosting and MLP. Experimentation framework was coded in MATLAB and Java. A custom and reusable object-oriented database was created using ObjectDB and interfaced with experimentation framework. The MATLAB interface to SVM$^{light}$ was obtained from http://www.igi.tugraz.at/aschwaig/software.html.
**Example 7: Assignment of identities to markers identified by SELDI-ToF/MS**

In order to assign identities to the protein biomarkers identified by SELDI-ToF/MS as being capable of discriminating sera from patients with Tuberculosis from sera from normal individuals, a pool of sera from 20 patients with TB and a second pool of sera from 20 healthy controls were generated. These were separated by 2D gel electrophoresis. To match the SELDI peak mass of a biomarker to the mass of a protein spot within the 2D gel, a second 2D gel was run where each spot was excised and the protein eluted passively from it to generate a solution of the full length protein. The solution of full length protein was analysed by SELDI-ToF/MS to generate a spectrum with a single peak. This mass was then compared with the original SELDI-ToF/MS biomarker mass list. A match between the two SELDI-ToF masses identifies the gel spot as the one corresponding to the SELDI-ToF/MS biomarker peak.

The gel spots from the matching 2D gel were removed and in-gel digested with trypsin to produce a peptide mixture diagnostic for that protein. This mixture was then analysed by LC/MS/MS to give a high probability prediction of identity based upon a BLAST search of the genome database.

Three biomarkers have been definitively identified in this way as shown in Table 5. The TB marker having an m/z value of 18394 is a serum albumin precursor, the TB marker having an m/z value of 11454 is Apo-A1 and the TB marker having an m/z value of 13774 is transthyretin.

**Example 8: Identification of further markers**

Analysis of the 2D gels containing serum proteins from TB patients and control subjects revealed that some proteins which did not appear to correspond to the markers identified by SELDI-ToF were differentially present in TB sera and sera from control subjects. The proteins were identified by removing the protein spots and in-gel digestion with trypsin to produce a peptide mixture diagnostic for that protein. The mixture was then analysed by LC/MS/MS to give a high probability prediction of identity based upon a BLAST search of the genome database. The additional markers identified were apolipoprotein-A2, hemoglobin beta, haptoglobin
protein, DEP domain protein, leucine-rich-alpha-2-glycoprotein (A2GL or LRG1) and hypothetical protein DFKZp6671032.

The results of this analysis are shown in Table 6. As can be seen from Table 6, transthyretin was identified from both the control gel and the TB gel. However, transthyretin was expressed at a lower level in the TB gel compared to the control gel, confirming that transthyretin is a negative marker of TB. Similarly, Apo-A2 expression is lower in the TB gel compared to the control gel and so Apo-A2 is negative marker of TB. Similarly, haptoglobin and hemoglobin beta are both expressed at a lower level in the TB gel compared to the control gel and so are negative markers of TB. A2GL (LRG1) and DEP domain protein, on the other hand, are upregulated in the TB gel compared to the control gel and so are positive markers of TB.

Hypothetical protein DFKZp6671032 was found only in the control gel and so is a negative marker of TB.

REFERENCES


Koyanagi, A., Kuffo, D., Gresely, L., Shenkin, A. & Cuevas, L.E.


Table 1. Participant demographics

<table>
<thead>
<tr>
<th></th>
<th>TUBERCULOSIS¹</th>
<th></th>
<th></th>
<th>CONTROLS</th>
<th></th>
<th></th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td>Total</td>
<td>Train</td>
<td>Test</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Total no. of patients (%)²</td>
<td>102</td>
<td>77</td>
<td>179</td>
<td>91</td>
<td>79</td>
<td>170</td>
<td>349</td>
</tr>
<tr>
<td>Age (years) [mean (range)]</td>
<td>31 (16-86)</td>
<td>35 (19-84)</td>
<td>32 (16-86)</td>
<td>44 (16-88)</td>
<td>46 (14-84)</td>
<td>45 (16-84)</td>
<td>206 (14-88)</td>
</tr>
<tr>
<td>Ethnic Origin (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-Saharan African</td>
<td>81 (79.4)</td>
<td>60 (77.9)</td>
<td>141 (78.8)</td>
<td>28 (30.7)</td>
<td>21 (26.5)</td>
<td>49 (28.8)</td>
<td>110</td>
</tr>
<tr>
<td>African not specified</td>
<td>3 (2.9)</td>
<td>1 (1.3)</td>
<td>4 (2.2)</td>
<td>3 (3.3)</td>
<td>3 (3.8)</td>
<td>6 (3.5)</td>
<td>90</td>
</tr>
<tr>
<td>Asian</td>
<td>13 (12.7)</td>
<td>9 (11.6)</td>
<td>22 (12.3)</td>
<td>3 (3.3)</td>
<td>0</td>
<td>3 (1.7)</td>
<td>25</td>
</tr>
<tr>
<td>White Caucasian</td>
<td>5 (4.9)</td>
<td>7 (9)</td>
<td>12 (6.7)</td>
<td>35 (38.4)</td>
<td>29 (36.7)</td>
<td>64 (37.6)</td>
<td>76</td>
</tr>
<tr>
<td>Not recorded</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>26</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>Collection Site:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uganda</td>
<td>80 (78.4)</td>
<td>59 (76.6)</td>
<td>139 (77.6)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>139</td>
</tr>
<tr>
<td>The Gambia</td>
<td>1 (0.9)</td>
<td>1 (1.3)</td>
<td>2 (1.1)</td>
<td>11 (12)</td>
<td>10 (12.6)</td>
<td>21 (12.3)</td>
<td>23</td>
</tr>
<tr>
<td>Angola</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10 (10.9)</td>
<td>9 (11.3)</td>
<td>19 (11.1)</td>
<td>19</td>
</tr>
<tr>
<td>UK (SGH)</td>
<td>21 (20.5)</td>
<td>17 (22)</td>
<td>38 (21.2)</td>
<td>70 (76.9)</td>
<td>60 (75.9)</td>
<td>130 (76.4)</td>
<td>168</td>
</tr>
<tr>
<td>HIV serology:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV positive (%)</td>
<td>35 (34.3)</td>
<td>24 (31.1)</td>
<td>59 (32.9)</td>
<td>2 (2.2)</td>
<td>3 (3.8)</td>
<td>5 (2.9)</td>
<td>64</td>
</tr>
<tr>
<td>CD4 count ≥000x10⁶/μl (%)³</td>
<td>19 (54.3)</td>
<td>13 (54.2)</td>
<td>32 (54.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD4 count &lt; 200x10⁶/μl (%)</td>
<td>15 (42.8)</td>
<td>11 (45.8)</td>
<td>26 (44.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIV negative (%)</td>
<td>60 (58.8)</td>
<td>45 (58.4)</td>
<td>105 (58.6)</td>
<td>12 (13.2)</td>
<td>8 (10.1)</td>
<td>20 (11.8)</td>
<td>125</td>
</tr>
<tr>
<td>HIV not determined (%)</td>
<td>7 (6.8)</td>
<td>8 (10.3)</td>
<td>15 (8.3)</td>
<td>77 (84.6)</td>
<td>68 (86)</td>
<td>145 (85.2)</td>
<td>160</td>
</tr>
</tbody>
</table>

¹ 12 TB patients had received between 1 and 7 days of chemotherapy at time of recruitment to the study.
² Demographic data were missing for 24 patients in the training set and 25 in the testing set.
³ CD4 counts were available for HIV seropositive patients; there was no value available for 6 seropositive patients.
Table 2. Characteristics of TB and control subjects

<table>
<thead>
<tr>
<th>a. TB patient characteristics</th>
<th>Train</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptomatic (%)</td>
<td>100 (98)</td>
<td>74 (96.1)</td>
<td>174 (97.2)</td>
</tr>
<tr>
<td>Persistent Cough</td>
<td>98 (96)</td>
<td>74 (96.1)</td>
<td>171 (95.5)</td>
</tr>
<tr>
<td>Haemoptysis</td>
<td>5 (4.9)</td>
<td>1 (1.3)</td>
<td>6 (3.3)</td>
</tr>
<tr>
<td>Night sweats/fever</td>
<td>68 (66.6)</td>
<td>53 (66.8)</td>
<td>121 (67.6)</td>
</tr>
<tr>
<td>Weight loss (≥5% ±&lt;5%)</td>
<td>86 (84.3)</td>
<td>60 (77.9)</td>
<td>146 (81.5)</td>
</tr>
<tr>
<td>Symptom duration pre-sampling (mean [range])</td>
<td>122.6 [13-449]</td>
<td>129.5 [12-754]</td>
<td>126 (12-754)</td>
</tr>
<tr>
<td>Smear Positive</td>
<td>89 (87.2)</td>
<td>66 (85.7)</td>
<td>155 (86.5)</td>
</tr>
<tr>
<td>Pulmonary disease</td>
<td>77 (75.4)</td>
<td>64 (83.1)</td>
<td>141 (78.7)</td>
</tr>
<tr>
<td>Extra-pulmonary disease</td>
<td>2 (1.9)</td>
<td>2 (2.6)</td>
<td>4 (2.2)</td>
</tr>
<tr>
<td>Pulmonary and extra-pulmonary</td>
<td>22 (21.5)</td>
<td>11 (14.2)</td>
<td>33 (18.4)</td>
</tr>
<tr>
<td>Abnormal CXR (%)</td>
<td>95 (93.1)</td>
<td>67 (87)</td>
<td>162 (90.5)</td>
</tr>
<tr>
<td>Cavitary Disease (%)</td>
<td>66 (64.7)</td>
<td>49 (63.6)</td>
<td>115 (64.2)</td>
</tr>
<tr>
<td>Previous BCG vaccination (%)</td>
<td>36 (35.3)</td>
<td>26 (33.8)</td>
<td>62 (34.6)</td>
</tr>
<tr>
<td>Skin test positive (%)</td>
<td>56 (54.9)</td>
<td>36 (46.8)</td>
<td>92 (51.4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b. Control diagnostic groups3</th>
<th>Train</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inflammatory bowel disease</td>
<td>10 (10.9)</td>
<td>6 (7.5)</td>
<td>16 (9.4)</td>
</tr>
<tr>
<td>Sarcoidosis</td>
<td>6 (6.5)</td>
<td>7 (8.8)</td>
<td>13 (7.6)</td>
</tr>
<tr>
<td>Respiratory infections</td>
<td>27 (29.6)</td>
<td>24 (30.3)</td>
<td>51 (30)</td>
</tr>
<tr>
<td>Other Infections:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malaria (P. falciparum)</td>
<td>4 (4.4)</td>
<td>3 (3.8)</td>
<td>7 (4.1)</td>
</tr>
<tr>
<td>HAT (T. b. gambiense)</td>
<td>10 (10.9)</td>
<td>9 (11.3)</td>
<td>19 (11.1)</td>
</tr>
<tr>
<td>Others6</td>
<td>1 (1.1)</td>
<td>2 (2.5)</td>
<td>3 (1.7)</td>
</tr>
<tr>
<td>Neurological disease6</td>
<td>13 (14.2)</td>
<td>13 (16.4)</td>
<td>26 (15.2)</td>
</tr>
<tr>
<td>Autoimmune disease6</td>
<td>6 (6.5)</td>
<td>3 (3.8)</td>
<td>9 (5.2)</td>
</tr>
<tr>
<td>Myeloma/monoclonal gammopathy</td>
<td>2 (2.2)</td>
<td>3 (3.8)</td>
<td>5 (2.9)</td>
</tr>
<tr>
<td>Healthy volunteers</td>
<td>12 (13.1)</td>
<td>9 (11.3)</td>
<td>21 (12.3)</td>
</tr>
</tbody>
</table>

1 Definite history of BCG vaccination and/or presence of scar. Data missing from 38 patients.
2 Mantoux reaction ≥ 5 mm greatest diameter of induration or Heaf grade ≥ 2. Data missing from 46 patients.
3 12 control subjects were taking high dose systemic steroids (prednisolone ≥ 60mg/day or dexamethasone ≥ 2mg/day).
49 patients with HAT had advanced (neurological disease) based on detection of parasites and/or >5 white cells/mm³ in CSF.
5 Visceral leishmaniasis (1), meningococcal septicaemia (1), staphylococcal cellulitis (1).
6 Cerebral neoplasia (12), cerebral abscess in association with infective endocarditis (1), myasthenia gravis (2), multiple sclerosis (5) and lumbar disc prolapse (6).
7 Rheumatoid arthritis (3), systemic lupus erythematosus (4), systemic sclerosis (1), overlap syndrome (1).
Table 3. Diagnostic Performance of classifiers

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Output</th>
<th>Actual</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Support Vector Machine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel: Gaussian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sigma = 0.00004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soft Margin = 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Support Vector Machine</strong></td>
<td>TB 72</td>
<td>4</td>
<td>94.23</td>
<td>93.50</td>
<td>94.93</td>
</tr>
<tr>
<td><strong>ADTree + AdaBoost</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100 iterations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight threshold = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ADTree + AdaBoost</strong></td>
<td>TB 72</td>
<td>7</td>
<td>92.30</td>
<td>93.50</td>
<td>91.13</td>
</tr>
<tr>
<td><strong>C4.5 Tree + AdaBoost</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100 iterations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight threshold = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>C4.5 Tree + AdaBoost</strong></td>
<td>TB 71</td>
<td>8</td>
<td>91.02</td>
<td>92.20</td>
<td>89.87</td>
</tr>
<tr>
<td><strong>Tree Classifier C5.0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boost = 10, Global Pruning 25%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tree Classifier C5.0</strong></td>
<td>TB 72</td>
<td>10</td>
<td>90.38</td>
<td>93.51</td>
<td>87.34</td>
</tr>
<tr>
<td><strong>Support Vector Machine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel: polynomial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dimension = 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soft Margin = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Support Vector Machine</strong></td>
<td>TB 71</td>
<td>9</td>
<td>88.46</td>
<td>92.20</td>
<td>84.81</td>
</tr>
<tr>
<td><strong>SLP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized Shuffled Presentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SLP</strong></td>
<td>TB 68</td>
<td>12</td>
<td>86.54</td>
<td>88.31</td>
<td>84.81</td>
</tr>
<tr>
<td><strong>MLP [1 HL (111 N)]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning rate = 0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Momentum = 0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized 500 epochs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MLP [1 HL (111 N)]</strong></td>
<td>TB 65</td>
<td>9</td>
<td>86.53</td>
<td>84.41</td>
<td>88.60</td>
</tr>
</tbody>
</table>

TB = tuberculosis; C = controls. ADTree = adaptive decision tree. AdaBoost = adaptive boosting. SLP = single layer perceptron. MLP = multi layered perceptron. HL = hidden layers. N = neurons. Key in *italics* and colors corresponds to name of classifier in Fig 1a.
### Table 4: Classifiers performance on selected mass cluster peaks and biomarkers

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>TPR</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 positive correlated and 10 negative correlated</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>199 (remaining)</td>
<td>0.86</td>
<td>0.82</td>
<td>0.90</td>
<td>0.82</td>
<td>0.10</td>
</tr>
<tr>
<td>10 positive correlated</td>
<td>0.78</td>
<td>0.75</td>
<td>0.80</td>
<td>0.75</td>
<td>0.20</td>
</tr>
<tr>
<td>209 (remaining)</td>
<td>0.89</td>
<td>0.83</td>
<td>0.95</td>
<td>0.83</td>
<td>0.05</td>
</tr>
<tr>
<td>10 negative correlated</td>
<td>0.85</td>
<td>0.88</td>
<td>0.81</td>
<td>0.88</td>
<td>0.19</td>
</tr>
<tr>
<td>209 (remaining)</td>
<td>0.89</td>
<td>0.87</td>
<td>0.91</td>
<td>0.87</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Markers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transthyretin</td>
<td>0.73</td>
<td>0.85</td>
<td>0.61</td>
<td>0.85</td>
<td>0.39</td>
</tr>
<tr>
<td>CRP</td>
<td>0.80</td>
<td>0.85</td>
<td>0.74</td>
<td>0.85</td>
<td>0.26</td>
</tr>
<tr>
<td>Neopterin</td>
<td>0.73</td>
<td>0.78</td>
<td>0.67</td>
<td>0.78</td>
<td>0.33</td>
</tr>
<tr>
<td>SAA</td>
<td>0.82</td>
<td>0.86</td>
<td>0.77</td>
<td>0.86</td>
<td>0.23</td>
</tr>
<tr>
<td>Neopterin - SAA</td>
<td>0.74</td>
<td>0.77</td>
<td>0.71</td>
<td>0.77</td>
<td>0.29</td>
</tr>
<tr>
<td>CRP - SAA</td>
<td>0.83</td>
<td>0.86</td>
<td>0.80</td>
<td>0.86</td>
<td>0.20</td>
</tr>
<tr>
<td>CRP - Neopterin</td>
<td>0.80</td>
<td>0.78</td>
<td>0.83</td>
<td>0.78</td>
<td>0.17</td>
</tr>
<tr>
<td>Transthyretin - SAA</td>
<td>0.81</td>
<td>0.92</td>
<td>0.70</td>
<td>0.92</td>
<td>0.30</td>
</tr>
<tr>
<td>Transthyretin - Neopterin</td>
<td>0.80</td>
<td>0.95</td>
<td>0.65</td>
<td>0.95</td>
<td>0.35</td>
</tr>
<tr>
<td>Transthyretin - CRP</td>
<td>0.82</td>
<td>0.92</td>
<td>0.71</td>
<td>0.92</td>
<td>0.29</td>
</tr>
<tr>
<td>Transthyretin - CRP - Neopterin</td>
<td>0.84</td>
<td>0.82</td>
<td>0.86</td>
<td>0.82</td>
<td>0.14</td>
</tr>
<tr>
<td>Transthyretin - CRP - SAA</td>
<td>0.82</td>
<td>0.92</td>
<td>0.72</td>
<td>0.92</td>
<td>0.28</td>
</tr>
<tr>
<td>Transthyretin - Neopterin - SAA</td>
<td>0.80</td>
<td>0.92</td>
<td>0.67</td>
<td>0.92</td>
<td>0.33</td>
</tr>
<tr>
<td>CRP - Neopterin - SAA</td>
<td>0.82</td>
<td>0.85</td>
<td>0.80</td>
<td>0.85</td>
<td>0.20</td>
</tr>
<tr>
<td>Transthyretin - CRP - Neopterin - SAA</td>
<td>0.79</td>
<td>0.89</td>
<td>0.68</td>
<td>0.89</td>
<td>0.32</td>
</tr>
</tbody>
</table>
### Table 5: Identification of Protein Markers

<table>
<thead>
<tr>
<th>SELDI-TOF/MS BIOMARKER</th>
<th>DATA DERIVED FROM 2D GELS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive in TB</strong></td>
<td></td>
</tr>
<tr>
<td>Mass</td>
<td>PE Mass</td>
</tr>
<tr>
<td>18394</td>
<td>18474</td>
</tr>
<tr>
<td>11720</td>
<td>11718</td>
</tr>
<tr>
<td>11454</td>
<td>11601</td>
</tr>
<tr>
<td>11506</td>
<td>11698</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Negative in TB</strong></td>
<td></td>
</tr>
<tr>
<td>SELDI-TOF/MS BIOMARKER</td>
<td>DATA DERIVED FROM 2D GELS</td>
</tr>
<tr>
<td>Mass</td>
<td>PE Mass</td>
</tr>
<tr>
<td>13774</td>
<td>13851</td>
</tr>
</tbody>
</table>

### Table 6: Protein Markers identified by 2D gel analysis

<table>
<thead>
<tr>
<th>Spots in TB gel</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE Mass (accurate)</td>
</tr>
<tr>
<td>8648</td>
</tr>
<tr>
<td>8771</td>
</tr>
<tr>
<td>16020</td>
</tr>
<tr>
<td>13876</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spots in Control gel</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE Mass (accurate)</td>
</tr>
<tr>
<td>13851</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>6.5, 5.9 and 6.3</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Bold** text denotes that the protein spot was more intense than the equivalent spot in the other gel.

*Italic* text denotes the protein spot was less intense than the equivalent spot in the other gel.
CLAIMS

1. A method of diagnosing tuberculosis (TB) in a test subject, said method comprising:

   (i) providing expression data of two or more markers in a subject, wherein at least two of said markers are selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp667I032; and

   (ii) comparing said expression data to expression data of said marker from a group of control subjects, wherein said control subjects comprise patients suffering from inflammatory conditions other than TB, thereby determining whether or not said test subject has TB.

2. A method according to claim 1, wherein said group of control subjects is selected from two or more of patients with respiratory infections, patients with sarcoidosis, patients with inflammatory bowel disease, patients with malaria, patients with human African trypanosomiasis (HAT), patients with neurological disease, patients with autoimmune disease, patients with myeloma and healthy subjects.

3. A method of diagnosing tuberculosis (TB), said method comprising:

   (i) providing expression data of two or more markers in a subject, wherein at least two of said markers are selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp667I032; and

   (ii) determining whether expression of said markers is indicative of TB

4. A method according to any one of the preceding claims, wherein one of said markers is transthyretin.

5. A method according to any one of the preceding claims, wherein said markers comprise transthyretin, CRP and neopterin.
6. A method according to any one of the preceding claims, wherein step (ii) is implemented using a computer system.

7. A method according to claim 6, wherein the computer system is programmed with a trained machine learning classifier.

8. A method according to claim 7, wherein said machine learning classifier is a support vector machine (SVM).

9. A method according to claim 3, wherein step (ii) comprises comparing expression of said markers in said subject to expression of said markers in a control subject.

10. A method according to claim 9, wherein the control subject is a patient suffering from an inflammatory condition other than TB.

11. A method according to claim 9 or 10, wherein said control subjects are selected from one or more of patients with respiratory infections, patients with sarcoidosis, patients with inflammatory bowel disease, patients with malaria, patients with human African trypanosomiasis (HAT), patients with neurological disease, patients with autoimmune disease, patients with myeloma and healthy subjects.

12. A method according to any one of the preceding claims, wherein step (ii) comprises comparing expression of said markers in said subject to expression of said markers in a TB patient.

13. A method according to claim 12, wherein said TB patient has been diagnosed as having TB by culture of Mycobacterium tuberculosis.

14. A method according to claim 12 or 13, wherein one or more patient having TB and/or one or more control subject is HIV positive.

15. A method according to any one of the preceding claims, wherein said markers comprise two or more of transthyretin, neopterin, CRP, SAA, serum albumin and Apo-A1 and one or more of apolipoprotein-A2, hemoglobin beta, haptoglobin protein, DEP domain protein, A2GL and hypothetical protein DFKZp667l032.

16. A method according to any one of the preceding claims, wherein said expression data is obtained by capture of said markers on a surface and detection of the captured markers.
17. A method according to claim 16, wherein said surface is a surface enhanced laser desorption and ionization (SELDI) probe and said detection is by SELDI-time of flight mass spectroscopy (SELDI-ToF MS).

18. A method according to claim 17, wherein said markers comprise one or more positively correlated markers having m/z values of about M18394_9, about M8952_75, about M11720_0, about M11454_1, about M18591_2, about M11488_1, about M9076_68, about M8895_13 and about M10856_8 and/or one or more negatively correlated markers having m/z values of about M4100_03, about M3898_52, about M13972_1, about M3322_01, about M2956_45, about M5644_96, about M3939_63, about M4056_39 and about M6649_74.

19. A method according to claim 18, wherein said markers comprise all said positively correlated markers and/or all said negatively correlated markers.

20. A method according to claim 16, wherein said surface comprises specific binding reagents for said markers and said detection is by immunoassay.

21. A computer-implemented method of diagnosing TB, said method comprising:

(i) inputting expression data of two or more markers in a subject; and

(ii) determining whether expression of said markers is indicative of TB using a computer system programmed with a trained support vector machine (SVM)

thereby diagnosing whether or not said patient has TB.

22. A method according to claim 21, wherein said SVM has been trained using data obtained from patients diagnosed as having TB by culture of Mycobacterium tuberculosis and from control subjects selected from one or more of patients with respiratory infections, patients with sarcoidosis, patients with inflammatory bowel disease, patients with malaria, patients with human African trypanosomiasis (HAT), patients with neurological disease, patients with autoimmune disease, patients with myeloma and healthy subjects.

23. A method of training a support vector machine (SVM) classifier to diagnose tuberculosis (TB), said method comprising:

(i) providing training data which comprises:
(a) training data relating to two or more markers in each of a first set of TB patients; and
(b) training data relating to said two or more markers in each of a first set of control subjects;

(ii) using a SVM to discriminate the training data of TB patients from the training data of control subjects; thereby training the SVM to diagnose TB.

24. A method according to claim 23, said method further comprising:

(iii) providing testing data which comprises:

(a) testing data relating to said two or more markers in each of a second set of TB patients; and
(b) testing data relating to said two or more markers in each of a second set of control subjects;

(iv) determining the ability of the SVM to correctly discriminate the testing data of TB patients from the testing data of control subjects.

25. A method according to claim 23 or 24, wherein said control subjects are selected from one or more of patients with respiratory infections, patients with sarcoidosis, patients with inflammatory bowel disease, patients with malaria, patients with human African trypanosomiasis (HAT), patients with neurological disease, patients with autoimmune disease, patients with myeloma and healthy subjects.

26. A method according to any one of claims 23 to 25, wherein said training data and said testing data are obtained by SELDI analysis.

27. A method according to any one of claims 23 to 25, wherein said training and said testing data are obtained by immunoassay analysis.

28. A method according to any one of claims 23 to 27, wherein at least one of said markers is selected from CRP, neopterin, SAA, transthyretin, serum albumin and Apo-A1.

29. A method according to claim 28, wherein said markers comprise CRP, transthyretin and neopterin.

30. A method according to any one of claims 23 to 29, wherein at least one of said markers is selected from Apo-A2, hemoglobin beta, haptoglobin protein, DEP domain protein, A2GL and hypothetical protein DFKZp6671032.
31. An apparatus arranged to perform a method according to any one of claims 21 to 30 comprising:
   (i) means for receiving expression data of two or more markers in a sample from a subject;
   (ii) a module for determining whether said data is indicative of TB, wherein said module comprises a trained machine learning classifier capable of distinguishing data from a TB patient from data from a control subject; and
   (iii) means for indicating the results of said determination.

32. An apparatus according to claim 31, which is a personal computer.

33. A computer program executable by a computer system, the computer program being capable, on execution by the computer system, of causing the computer system to perform a method according to any one of claims 21 to 30.

34. A storage medium storing in a form readable by a computer system having a computer program according to claim 33.

35. A kit for diagnosing TB comprising:
   (i) means for detecting two or more markers; and
   (ii) a storage medium according to claim 34.

36. A kit for diagnosing TB comprising:
   (i) means for detecting two or more markers;
   (ii) instructions for inputting data relating to detection of said markers into an apparatus according to claim 31 or 32.

37. A kit according to claim 35 or 36, wherein said markers are selected from transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2, hemoglobin beta, haptoglobin protein, DEP domain protein, A2GL and hypothetical protein DFKZp667I032.

38. A kit for diagnosing TB comprising:
   (i) means for detecting two or more markers selected from transthyretin, neopterin, C-reactive protein (CRP), serum amyloid A (SAA), serum albumin, apolipoprotein-A1 (Apo-A1), apolipoprotein-A2 (Apo-A2), hemoglobin beta, haptoglobin protein, DEP domain protein, leucine-rich alpha-2-glycoprotein (A2GL) and hypothetical protein DFKZp667I032.
39. A kit according to any one of claims 35 to 38, wherein said means of
detecting two or more markers comprises a capture surface.

40. A kit according to claim 39, wherein said capture surface is a protein
chip.

41. A kit according to claim 39 or 40, wherein said capture surface
comprises specific binding reagents for said markers.

42. A kit according to claim 41, wherein said specific binding reagents
are antibodies or antibody fragments.

43. A kit according to any one of claims 37 to 42, wherein said markers
are transthyretin, neopterin and CRP.

44. A method according to any one of claims 1 to 30 further comprising
administering to a patient diagnosed as having TB, a medicament for treatment of
TB.

45. A method of identifying an agent for the treatment of TB, said method
comprising:

(i) contacting a test agent with transthyretin, neopterin, CRP, SAA,
serum albumin, Apo-A1, Apo-A2, hemoglobin beta, haptoglobin, DEP
domain protein or A2GL; and

(ii) determining whether test agent modulates the activity of said
transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2,
hemoglobin beta, haptoglobin, DEP domain protein or A2GL

thereby determining whether or not said test agent is suitable for use in the treatment
of TB.

46. A method of identifying an agent for the treatment of TB, said method
comprising:

(i) contacting cells 

ex vivo or in vivo 

with Mycobacterium tuberculosis 

and a test agent;

(iii) monitoring expression of one or more TB markers selected from
transthyretin, neopterin, CRP, SAA, serum albumin, Apo-A1, Apo-A2,
hemoglobin beta, haptoglobin, DEP domain protein and A2GL; and

(iv) determining whether test agent modulates the expression of said one
or more test markers,
thereby determining whether or not said test agent is suitable for use in the treatment of TB.
FIG. 1

D1

Training data from TB patients

Training data from control subjects

S1

Map TB data and control data to feature space

S2

Discriminate between TB data and control data to produce classifier

FIG. 2

D2

Testing data from TB patients

Testing data from control subjects

S3

Map TB data and control data to feature space

S4

Classify data as being TB data or control data using classifier

Is data classified correctly?
FIG. 3

D3

Data from test subject

↓

S5

Map data to feature space

↓

S6

Classify data

↓

Does subject have TB
<table>
<thead>
<tr>
<th>Met</th>
<th>Lys</th>
<th>Leu</th>
<th>Leu</th>
<th>Thr</th>
<th>Gly</th>
<th>Leu</th>
<th>Val</th>
<th>Phe</th>
<th>Cys</th>
<th>Ser</th>
<th>Leu</th>
<th>Val</th>
<th>Leu</th>
<th>Gly</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ser</th>
<th>Ser</th>
<th>Arg</th>
<th>Ser</th>
<th>Phe</th>
<th>Phe</th>
<th>Ser</th>
<th>Phe</th>
<th>Leu</th>
<th>Gly</th>
<th>Glu</th>
<th>Ala</th>
<th>Phe</th>
<th>Asp</th>
<th>Gly</th>
<th>Ala</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>25</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arg</th>
<th>Asp</th>
<th>Met</th>
<th>Trp</th>
<th>Arg</th>
<th>Ala</th>
<th>Tyr</th>
<th>Ser</th>
<th>Asp</th>
<th>Met</th>
<th>Arg</th>
<th>Glu</th>
<th>Ala</th>
<th>Asn</th>
<th>Tyr</th>
<th>Ile</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>40</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gly</th>
<th>Ser</th>
<th>Asp</th>
<th>Lys</th>
<th>Tyr</th>
<th>Phe</th>
<th>His</th>
<th>Ala</th>
<th>Arg</th>
<th>Gly</th>
<th>Asn</th>
<th>Tyr</th>
<th>Asp</th>
<th>Ala</th>
<th>Ala</th>
<th>Lys</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>55</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arg</th>
<th>Gly</th>
<th>Pro</th>
<th>Gly</th>
<th>Gly</th>
<th>Val</th>
<th>Trp</th>
<th>Ala</th>
<th>Ala</th>
<th>Glu</th>
<th>Ala</th>
<th>Ile</th>
<th>Ser</th>
<th>Asp</th>
<th>Ala</th>
<th>Arg</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>70</td>
<td>75</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Glu</th>
<th>Asn</th>
<th>Ile</th>
<th>Gln</th>
<th>Arg</th>
<th>Phe</th>
<th>Phe</th>
<th>Gly</th>
<th>His</th>
<th>Gly</th>
<th>Ala</th>
<th>Glu</th>
<th>Asp</th>
<th>Ser</th>
<th>Leu</th>
<th>Ala</th>
</tr>
</thead>
<tbody>
<tr>
<td>85</td>
<td>90</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Asp</th>
<th>Gln</th>
<th>Ala</th>
<th>Asn</th>
<th>Glu</th>
<th>Trp</th>
<th>Gly</th>
<th>Arg</th>
<th>Ser</th>
<th>Gly</th>
<th>Lys</th>
<th>Asp</th>
<th>Pro</th>
<th>Asn</th>
<th>His</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>105</td>
<td>110</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phe</th>
<th>Arg</th>
<th>Pro</th>
<th>Ala</th>
<th>Gly</th>
<th>Leu</th>
<th>Pro</th>
<th>Glu</th>
<th>Lys</th>
<th>Tyr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>115</td>
<td></td>
<td></td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Met</th>
<th>Glu</th>
<th>Lys</th>
<th>Leu</th>
<th>Leu</th>
<th>Cys</th>
<th>Phe</th>
<th>Leu</th>
<th>Val</th>
<th>Leu</th>
<th>Thr</th>
<th>Ser</th>
<th>Leu</th>
<th>Ser</th>
<th>His</th>
<th>Ala</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Phe | Gly | Gln | Thr | Asp | Met | Ser | Arg | Lys | Ala | Phe | Val | Phe | Pro | Lys | Glu |
Ser Asp Thr Ser Tyr Val Ser Leu Lys Ala Pro Leu Thr Lys Pro Leu
35  40  45
Lys Ala Phe Thr Val Cys Leu His Phe Tyr Thr Glu Leu Ser Ser Thr
50  55  60
Arg Gly Tyr Ser Ile Phe Ser Tyr Ala Thr Lys Arg Gln Asp Asn Glu
65  70  75  80
Ile Leu Ile Phe Trp Ser Lys Asp Ile Gly Tyr Ser Phe Thr Val Gly
85  90  95
Gly Ser Glu Ile Leu Phe Glu Val Pro Glu Val Thr Val Ala Pro Val
100 105 110
His Ile Cys Thr Ser Trp Glu Ser Ala Ser Gly Ile Val Glu Phe Trp
115 120 125
Val Asp Gly Lys Pro Arg Val Arg Lys Ser Leu Lys Gly Tyr Thr
130 135 140
Val Gly Ala Glu Ala Ser Ile Ile Leu Gly Glu Gln Glu Asp Ser Phe
145 150 155 160
Gly Gly Asn Phe Glu Gly Ser Gln Ser Leu Val Gly Asp Ile Gly Asn
165 170 175
Val Asn Met Trp Asp Phe Val Leu Ser Pro Asp Glu Ile Asn Thr Ile
180 185 190
Tyr Leu Gly Gly Pro Phe Ser Pro Asn Val Leu Asn Trp Arg Ala Leu
195 200 205
Lys Tyr Glu Val Gln Gly Glu Val Phe Thr Lys Pro Gln Leu Trp Pro
210 215 220

<210>  3
<211> 147
<212> PRT
<213> Homo sapiens

<400>  3
Met Ala Ser His Arg Leu Leu Leu Leu Cys Leu Ala Gly Leu Val Phe
1  5 10 15
Val Ser Glu Ala Gly Pro Thr Gly Thr Gly Glu Ser Lys Cys Pro Leu
20 25 30
Met Val Lys Val Leu Asp Ala Val Arg Gly Ser Pro Ala Ile Asn Val
35 40 45
Ala Val His Val Phe Arg Lys Ala Ala Asp Thr Trp Glu Pro Phe
50 55 60
Ala Ser Gly Lys Thr Ser Glu Ser Gly Glu Leu His Gly Leu Thr Thr
<table>
<thead>
<tr>
<th></th>
<th>65</th>
<th>70</th>
<th>75</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Glu Glu Glu Phe Val Glu Gly Ile Tyr Lys Val Glu Ile Asp Thr Lys</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>85</td>
<td>90</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ser Tyr Trp Lys Ala Leu Gly Ile Ser Pro Phe His Glu His Ala Glu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>105</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Val Val Phe Thr Ala Asn Asp Gly Pro Arg Arg Tyr Thr Ile Ala</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>115</td>
<td>120</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ala Leu Leu Ser Pro Tyr Ser Tyr Ser Thr Thr Ala Val Val Thr Asn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>135</td>
<td>140</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pro Lys Glu</td>
<td></td>
<td></td>
<td>145</td>
</tr>
<tr>
<td>&lt;210&gt;</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;211&gt;</td>
<td>609</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;212&gt;</td>
<td>PRT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;213&gt;</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;400&gt;</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Met Lys Trp Val Thr Phe Ile Ser Leu Leu Phe Leu Phe Ser Ser Ala</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Tyr Ser Arg Gly Val Phe Arg Arg Asp Ala His Lys Ser Glu Val Ala</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>25</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>His Arg Phe Lys Asp Leu Gly Glu Glu Asn Phe Lys Ala Leu Val Leu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>40</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ile Ala Phe Ala Gln Tyr Leu Gln Gln Cys Pro Phe Glu Asp His Val</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>55</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys Leu Val Asn Glu Val Thr Glu Phe Ala Lys Thr Cys Val Ala Asp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65</td>
<td>70</td>
<td>75</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Glu Ser Ala Glu Asn Cys Asp Lys Ser Leu His Thr Leu Phe Gly Asp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>85</td>
<td></td>
<td>90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lys Leu Cys Thr Val Ala Thr Leu Arg Glu Thr Tyr Gly Glu Met Ala</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>105</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asp Cys Ala Lys Gln Glu Pro Glu Ser Asn Glu Cys Phe Leu Gln</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>115</td>
<td></td>
<td>120</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td>His Lys Asp Asn Pro Asn Leu Pro Arg Leu Val Arg Pro Glu Val</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>135</td>
<td></td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>Asp Val Met Cys Thr Ala Phe His Asp Asn Glu Glu Thr Phe Leu Lys</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>145</td>
<td>150</td>
<td>155</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Lys Tyr Leu Tyr Glu Ile Ala Arg Arg His Pro Tyr Phe Tyr Ala Pro</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>165</td>
<td>170</td>
<td></td>
<td>175</td>
</tr>
</tbody>
</table>
Glu Leu Leu Phe Phe Ala Lys Arg Tyr Lys Ala Ala Phe Thr Glu Cys
180 185 190
Cys Gln Ala Ala Asp Lys Ala Ala Cys Leu Leu Pro Lys Leu Asp Glu
195 200 205
Leu Arg Asp Glu Gly Lys Ala Ser Ser Ala Lys Gln Arg Leu Lys Cys
210 215 220
Ala Ser Leu Gln Lys Phe Gly Glu Arg Ala Phe Lys Ala Trp Ala Val
225 230 235 240
Ala Arg Leu Ser Gln Arg Phe Pro Lys Ala Glu Phe Ala Glu Val Ser
245 250 255
Lys Leu Val Thr Asp Leu Thr Lys Val His Thr Glu Cys Cys His Gly
260 265 270
Asp Leu Leu Glu Cys Ala Asp Asp Arg Ala Asp Leu Ala Lys Tyr Ile
275 280 285
Cys Glu Asn Gln Asp Ser Ile Ser Ser Lys Leu Lys Glu Cys Cys Glu
290 295 300
Lys Pro Leu Leu Glu Lys Ser His Cys Ile Ala Glu Val Glu Asn Asp
305 310 315 320
Glu Met Pro Ala Asp Leu Pro Ser Leu Ala Ala Asp Phe Val Glu Ser
325 330 335
Lys Asp Val Cys Lys Asn Tyr Ala Glu Ala Lys Asp Val Phe Leu Gly
340 345 350
Met Phe Leu Tyr Glu Tyr Ala Arg Arg His Pro Asp Tyr Ser Val Val
355 360 365
Leu Leu Leu Arg Leu Ala Lys Thr Tyr Glu Thr Thr Leu Glu Lys Cys
370 375 380
Cys Ala Ala Ala Asp Pro His Glu Cys Tyr Ala Lys Val Phe Asp Glu
385 390 395 400
Phe Lys Pro Leu Val Glu Glu Pro Glu Asn Leu Ile Lys Gln Asn Cys
405 410 415
Glu Leu Phe Glu Glu Leu Gly Glu Tyr Lys Phe Glu Asn Ala Leu Leu
420 425 430
Val Arg Tyr Thr Lys Val Pro Glu Val Ser Thr Pro Thr Leu Val
435 440 445
Glu Val Ser Arg Asn Leu Gly Lys Val Gly Ser Lys Cys Cys His Lys
450 455 460
Pro Gly Ala Lys Arg Met Pro Cys Ala Glu Asp Tyr Leu Ser Val Val
465 470 475 480
Leu Asn Gln Leu Cys Val Leu His Glu Lys Thr Pro Val Ser Asp Arg
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val Thr Lys Cys Cys Thr Glu Ser Leu Val Asn Arg Arg Pro Cys Phe</td>
<td>1-35</td>
</tr>
<tr>
<td>Ser Ala Leu Glu Val Asp Glu Thr Tyr Val Pro Lys Glu Phe Asn Ala</td>
<td>36-50</td>
</tr>
<tr>
<td>Glu Thr Phe Thr Phe His Ala Asp Ile Cys Thr Leu Ser Glu Lys Glu</td>
<td>51-60</td>
</tr>
<tr>
<td>Arg Gln Ile Lys Lys Gln Thr Ala Val Leu Val Glu Leu Val Lys His Lys</td>
<td>61-70</td>
</tr>
<tr>
<td>Pro Lys Ala Thr Lys Glu Gln Leu Lys Ala Val Met Asp Asp Phe Ala</td>
<td>71-80</td>
</tr>
<tr>
<td>Ala Phe Val Glu Lys Cys Cys Lys Ala Asp Asp Lys Glu Thr Cys Phe</td>
<td>81-90</td>
</tr>
<tr>
<td>Ala Glu Glu Gly Lys Lys Leu Val Ala Ala Ser Gln Ala Ala Leu Gly</td>
<td>91-100</td>
</tr>
</tbody>
</table>

Leu

<210> 5
<211> 267
<212> PRT
<213> Homo sapiens

<400> 5

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met Lys Ala Ala Val Leu Thr Leu Ala Val Leu Phe Leu Thr Gly Ser</td>
<td>1-15</td>
</tr>
<tr>
<td>Gln Ala Arg His Phe Trp Gln Gln Asp Glu Pro Pro Gln Ser Pro Trp</td>
<td>16-30</td>
</tr>
<tr>
<td>Asp Arg Val Lys Asp Leu Ala Thr Val Tyr Val Asp Val Leu Lys Asp</td>
<td>31-45</td>
</tr>
<tr>
<td>Ser Gly Arg Asp Tyr Val Ser Gln Phe Glu Gly Ser Ala Leu Gly Lys</td>
<td>46-60</td>
</tr>
<tr>
<td>Gln Leu Asn Leu Lys Leu Leu Asn Trp Asp Ser Val Thr Ser Thr</td>
<td>61-75</td>
</tr>
<tr>
<td>Phe Ser Lys Leu Arg Glu Gln Leu Gly Pro Val Thr Gln Glu Phe Trp</td>
<td>76-90</td>
</tr>
<tr>
<td>Asp Asn Leu Glu Lys Glu Thr Glu Gly Leu Arg Gln Glu Met Ser Lys</td>
<td>91-110</td>
</tr>
<tr>
<td>Asp Leu Glu Glu Val Lys Ala Val Gln Pro Tyr Leu Asp Asp Phe</td>
<td>111-120</td>
</tr>
<tr>
<td>Gln Lys Lys Trp Gln Glu Glu Met Glu Leu Tyr Arg Gln Lys Val Glu</td>
<td>121-130</td>
</tr>
<tr>
<td>Sequence</td>
<td>Number</td>
</tr>
<tr>
<td>----------</td>
<td>--------</td>
</tr>
<tr>
<td>Pro Leu Arg Ala Glu Leu Gln Glu Gly Ala Arg Gln Lys Leu His Glu</td>
<td>130</td>
</tr>
<tr>
<td>Leu Gln Glu Lys Leu Ser Pro Leu Gly Glu Glu Met Arg Asp Arg Ala</td>
<td>135</td>
</tr>
<tr>
<td>Arg Ala His Val Asp Ala Leu Arg Thr His Leu Ala Pro Tyr Ser Asp</td>
<td>140</td>
</tr>
<tr>
<td>Glu Leu Arg Gln Arg Leu Ala Ala Arg Leu Glu Ala Leu Lys Glu Asn</td>
<td>145</td>
</tr>
<tr>
<td>Gly Gly Ala Arg Leu Ala Glu Tyr His Ala Lys Ala Thr Glu His Leu</td>
<td>150</td>
</tr>
<tr>
<td>Ser Thr Leu Ser Glu Ala Lys Pro Ala Leu Glu Asp Leu Arg Gln</td>
<td>155</td>
</tr>
<tr>
<td>Gly Leu Leu Pro Val Leu Glu Ser Phe Lys Val Ser Phe Leu Ser Ala</td>
<td>160</td>
</tr>
<tr>
<td>Leu Glu Glu Tyr Thr Lys Lys Leu Asn Thr Gln</td>
<td>165</td>
</tr>
</tbody>
</table>

<210> 6
<211> 347
<212> PRT
<213> Homo sapiens

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met Ser Ser Trp Ser Arg Gln Arg Pro Lys Ser Pro Gly Gly Ile Gln</td>
<td>1</td>
</tr>
<tr>
<td>Pro His Val Ser Arg Thr Leu Phe Leu Leu Leu Leu Ala Ala Ser</td>
<td>5</td>
</tr>
<tr>
<td>Ala Trp Gly Val Thr Leu Ser Pro Lys Asp Cys Gln Val Phe Arg Ser</td>
<td>10</td>
</tr>
<tr>
<td>Asp His Gly Ser Ser Ile Ser Cys Gln Pro Pro Ala Glu Ile Pro Gly</td>
<td>15</td>
</tr>
<tr>
<td>Tyr Leu Pro Ala Asp Thr Val His Leu Ala Val Glu Phe Phe Asn Leu</td>
<td>20</td>
</tr>
<tr>
<td>Thr His Leu Pro Ala Asn Leu Leu Gln Gly Ala Ser Lys Leu Gln Glu</td>
<td>25</td>
</tr>
<tr>
<td>Leu His Leu Ser Ser Asn Gly Leu Glu Ser Leu Ser Pro Glu Phe Leu</td>
<td>30</td>
</tr>
<tr>
<td>Arg Pro Val Pro Gln Leu Arg Val Leu Asp Leu Thr Arg Asn Ala Leu</td>
<td>35</td>
</tr>
<tr>
<td>Thr Gly Leu Pro Pro Gly Leu Phe Gln Ala Ser Ala Thr Leu Asp Thr</td>
<td>40</td>
</tr>
</tbody>
</table>
Leu Val Leu Lys Glu Asn Gln Leu Glu Val Leu Glu Val Ser Trp Leu
130 135 140
| 145 | 150 | 155 | 160 |
His Gly Leu Lys Ala Leu Gly His Leu Asp Leu Ser Gly Asn Arg Leu
165 170 175
Arg Lys Leu Pro Pro Gly Leu Leu Ala Asn Phe Thr Leu Leu Arg Thr
180 185 190
Leu Asp Leu Gly Glu Asn Gln Leu Glu Thr Leu Pro Pro Asp Leu Leu
195 200 205
| 210 | 215 | 220 |
Arg Gly Pro Leu Gln Leu Glu Arg Leu His Leu Glu Gly Asn Lys Leu
225
| 230 | 235 | 240 |
Gln Val Leu Gly Lys Asp Leu Leu Leu Pro Gln Pro Asp Leu Arg Tyr
255
| 260 | 265 | 270 |
Leu Phe Leu Asn Gly Asn Lys Leu Ala Arg Val Ala Ala Gly Ala Phe
275
| 280 | 285 |
Gln Gly Leu Arg Gln Leu Asp Met Leu Asp Leu Ser Asn Asn Ser Leu
290
| 295 | 300 |
Asp Met Arg Asp Gly Phe Asp Ile Ser Gly Asn Pro Trp Ile Cys Asp
305
| 310 | 315 | 320 |
Gln Asn Leu Ser Asp Leu Tyr Arg Trp Leu Gln Ala Gln Lys Asp Lys
325
| 330 | 335 |
Met Phe Ser Gln Asn Thr Arg Cys Ala Gly Pro Glu Ala Val Lys
Gly Gln Thr Leu Leu Ala Val Ala Lys Ser Gln
| 340 | 345 |

<210> 7
<211> 105
<212> PRT
<213> Homo sapiens

<400> 7
Met Val His Leu Thr Pro Glu Glu Lys Ser Ala Val Thr Ala Leu Trp
1 5 10 15
Gly Lys Val Asn Val Asp Ala Val Gly Gly Glu Ala Leu Gly Arg Leu
20 25 30
Leu Val Val Tyr Pro Trp Thr Gln Arg Phe Phe Glu Ser Phe Gly Asp
35 40 45
Leu Ser Thr Pro Asp Ala Val Met Gly Asn Pro Lys Val Lys Ala His
Val Gly Lys Lys Gln Leu Val Glu Ile Glu Lys Val Val Leu His Pro
225 230 235 240
Asn Tyr Ser Gln Val Asp Ile Gly Leu Ile Lys Leu Lys Gln Lys Val
245 250 255
Ser Val Asn Glu Arg Val Met Pro Ile Cys Leu Pro Ser Lys Asp Tyr
260 265 270
Ala Glu Val Gly Arg Val Gly Tyr Val Ser Gly Trp Gly Arg Asn Ala
275 280 285 290 295 300
Asn Phe Lys Phe Thr Asp His Leu Lys Tyr Val Met Leu Pro Val Ala
305 310 315 320
Asp Gln Asp Gln Cys Ile Arg His Tyr Glu Gly Ser Thr Val Pro Glu
325 330 335
Lys Lys Thr Pro Lys Ser Pro Val Gly Val Gln Pro Ile Leu Asn Glu
340 345 350
His Thr Phe Cys Ala Gly Met Ser Lys Tyr Gln Glu Asp Thr Cys Tyr
355 360 365
Gly Asp Ala Gly Ser Ala Phe Ala Val His Asp Leu Glu Glu Asp Thr
370 375 380
Trp Tyr Ala Thr Gly Ile Leu Ser Phe Asp Lys Ser Cys Ala Val Ala
385 390 395 400
Lys Thr Ile Ala Glu Asn
405

<210> 9
<211> 100
<212> PRT
<213> Homo sapiens
<400> 9

Met Lys Leu Leu Ala Ala Thr Val Leu Leu Leu Thr Ile Cys Ser Leu
1 5 10 15
Glu Gly Ala Leu Val Arg Arg Gln Ala Lys Glu Pro Cys Val Glu Ser
20 25 30
Leu Val Ser Gln Tyr Phe Gln Thr Val Thr Asp Tyr Gly Lys Asp Leu
35 40 45
Met Glu Lys Val Lys Ser Pro Glu Leu Gln Ala Glu Ala Lys Ser Tyr
50 55 60
Phe Glu Lys Ser Lys Glu Gln Leu Thr Pro Leu Ile Lys Lys Ala Gly
<table>
<thead>
<tr>
<th></th>
<th>65</th>
<th>70</th>
<th>75</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thr</td>
<td>Glu</td>
<td>Leu</td>
<td>Val</td>
<td>Asn</td>
</tr>
<tr>
<td>Pro</td>
<td>Ala</td>
<td>Thr</td>
<td>Gln</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;210&gt;</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;211&gt;</td>
<td>1572</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;212&gt;</td>
<td>PRT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;213&gt;</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Met</td>
<td>Arg</td>
<td>Thr</td>
<td>Thr</td>
<td>Lys</td>
</tr>
<tr>
<td>Gly</td>
<td>Gly</td>
<td>Ser</td>
<td>Asp</td>
<td>Asp</td>
</tr>
<tr>
<td>Ile</td>
<td>Lys</td>
<td>Leu</td>
<td>Gly</td>
<td>Asp</td>
</tr>
<tr>
<td>Ser</td>
<td>Pro</td>
<td>Leu</td>
<td>Leu</td>
<td>Leu</td>
</tr>
<tr>
<td>Glu</td>
<td>Thr</td>
<td>Ile</td>
<td>Ser</td>
<td>Val</td>
</tr>
<tr>
<td>Pro</td>
<td>Tyr</td>
<td>Gln</td>
<td>Asp</td>
<td>Val</td>
</tr>
<tr>
<td>Leu</td>
<td>Asp</td>
<td>Leu</td>
<td>Val</td>
<td>Thr</td>
</tr>
<tr>
<td>Asp</td>
<td>Met</td>
<td>Trp</td>
<td>Arg</td>
<td>Leu</td>
</tr>
<tr>
<td>Thr</td>
<td>Gln</td>
<td>Lys</td>
<td>Val</td>
<td>Phe</td>
</tr>
<tr>
<td>Trp</td>
<td>Val</td>
<td>Lys</td>
<td>Asn</td>
<td>Glu</td>
</tr>
<tr>
<td>Arg</td>
<td>Val</td>
<td>Phe</td>
<td>Arg</td>
<td>Ser</td>
</tr>
<tr>
<td>Met</td>
<td>Ser</td>
<td>Cys</td>
<td>Glu</td>
<td>Met</td>
</tr>
<tr>
<td>Glu</td>
<td>Lys</td>
<td>Ala</td>
<td>Val</td>
<td>Asn</td>
</tr>
<tr>
<td>Glu</td>
<td>Lys</td>
<td>Asn</td>
<td>Cys</td>
<td>Ser</td>
</tr>
<tr>
<td>Phe</td>
<td>Tyr</td>
<td>Asp</td>
<td>Ala</td>
<td>Lys</td>
</tr>
</tbody>
</table>

Phe Tyr Asp Ala Lys Ser Val Asp Glu Phe Pro Glu Ile Asn Arg Ala
Ser Ile Arg Gln Asp His Lys Gly Arg Phe Tyr Glu Asp Phe Tyr Lys  
245 250 255
Val Val Val Gln Asn Glu Arg Arg Glu Glu Trp Thr Ser Leu Leu Val  
260 265 270
Thr Ile Lys Leu Phe Ile Gln Tyr Pro Val Leu Val Arg Leu Glu  
275 280 285
Gln Ala Glu Gly Phe Pro Gln Gly Asp Asn Ser Thr Ser Ala Gln Gly  
290 295 300
Asn Tyr Leu Glu Ala Ile Asn Leu Ser Phe Asn Val Phe Asp Lys His  
305 310 315 320
Tyr Ile Asn Arg Asn Phe Asp Arg Thr Gly Gln Met Ser Val Val Ile  
325 330 335
Thr Pro Gly Val Gly Val Phe Glu Val Asp Arg Leu Leu Met Ile Leu  
340 345 350
Thr Lys Gln Arg Met Ile Asp Asn Gly Ile Gly Val Asp Leu Val Cys  
355 360 365
Met Gly Glu Gln Pro Leu His Ala Val Pro Leu Phe Lys Leu His Asn  
370 375 380 385
Arg Ser Ala Pro Arg Asp Ser Arg Leu Gly Asp Asp Tyr Asn Ile Pro  
390 395 400
His Trp Ile Asn His Ser Phe Tyr Thr Ser Lys Ser Gln Leu Phe Cys  
405 410 415
Asn Ser Phe Thr Pro Arg Ile Lys Leu Ala Gly Lys Lys Pro Ala Ser  
420 425 430
Glu Lys Ala Lys Asn Gly Arg Asp Thr Ser Leu Gly Ser Pro Lys Glu  
435 440 445
Ser Glu Asn Ala Leu Pro Ile Gln Val Asp Tyr Asp Ala Tyr Asp Ala  
450 455 460
Gln Val Phe Arg Leu Pro Gly Pro Ser Arg Ala Gln Cys Leu Thr Thr  
465 470 475 480
Cys Arg Ser Val Arg Glu Arg Glu Ser His Ser Arg Lys Ser Ala Ser  
485 490 495
Ser Cys Asp Val Ser Ser Ser Pro Ser Leu Pro Ser Arg Thr Leu Pro  
500 505 510
Thr Glu Glu Val Arg Ser Gln Ala Ser Asp Ser Ser Leu Gly Lys  
515 520 525
Ser Ala Asn Ile Leu Met Ile Pro His Pro His Leu His Gln Tyr Glu  
530 535 540
<table>
<thead>
<tr>
<th>Val Ser Ser Ser Leu Gly Tyr Thr Ser Thr Arg Asp Val Leu Glu Asn</th>
<th>545 550 555 560</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met Met Glu Pro Pro Gln Arg Asp Ser Ser Ala Pro Gly Arg Phe His</td>
<td>565 570 575</td>
</tr>
<tr>
<td>Val Gly Ser Ala Glu Ser Met Leu His Val Arg Pro Gly Gly Tyr Thr</td>
<td>580 585 590</td>
</tr>
<tr>
<td>Pro Gln Arg Ala Leu Ile Asn Pro Phe Ala Pro Ser Arg Met Pro Met</td>
<td>595 600 605</td>
</tr>
<tr>
<td>Lys Leu Thr Ser Asn Arg Arg Arg Trp Met His Thr Phe Pro Val Gly</td>
<td>610 615 620</td>
</tr>
<tr>
<td>Pro Ser Gly Glu Ala Ile Gln Ile His Glu Thr Arg Gln Asn Met</td>
<td>625 630 635 640</td>
</tr>
<tr>
<td>Ala Glu Leu Gln Gly Ser Gly Gln Arg Asp Pro Thr His Ser Ser Ala</td>
<td>645 650 655</td>
</tr>
<tr>
<td>Glu Leu Leu Glu Ala Tyr His Glu Ala Ala Gly Arg His Ser Asn</td>
<td>660 665 670</td>
</tr>
<tr>
<td>Ser Arg Gln Pro Gly Asp Gly Met Ser Phe Leu Asn Phe Ser Gly Thr</td>
<td>675 680 685</td>
</tr>
<tr>
<td>Glu Glu Leu Ser Val Gly Leu Leu Ser Asn Ser Gly Ala Gly Met Asn</td>
<td>690 695 700</td>
</tr>
<tr>
<td>Pro Arg Thr Gln Asn Lys Asp Ser Leu Glu Asp Ser Val Ser Thr Ser</td>
<td>705 710 715 720</td>
</tr>
<tr>
<td>Pro Asp Pro Met Pro Gly Phe Cys Cys Thr Val Gly Val Asp Trp Lys</td>
<td>725 730 735</td>
</tr>
<tr>
<td>Ser Leu Thr Thr Pro Ala Cys Leu Pro Leu Thr Thr Asp Tyr Phe Pro</td>
<td>740 745 750</td>
</tr>
<tr>
<td>Asp Arg Gln Gly Leu Gln Asn Asp Tyr Thr Glu Gly Cys Tyr Asp Leu</td>
<td>755 760 765</td>
</tr>
<tr>
<td>Leu Pro Glu Ala Asp Ile Asp Arg Arg Asp Glu Gly Val Gln Met</td>
<td>770 775 780</td>
</tr>
<tr>
<td>Thr Ala Gln Gln Val Phe Glu Glu Ile Cys Glu Arg Leu Met Gln</td>
<td>785 790 795 800</td>
</tr>
<tr>
<td>Gly Tyr Gln Ile Ile Val Gln Pro Lys Thr Gln Lys Pro Asn Pro Ala</td>
<td>805 810 815</td>
</tr>
<tr>
<td>Val Pro Pro Pro Leu Ser Ser Ser Pro Leu Tyr Ser Arg Gly Leu Val</td>
<td>820 825 830</td>
</tr>
<tr>
<td>Ser Arg Asn Arg Pro Glu Glu Gly Asp Gln Tyr Trp Leu Ser Met Gly</td>
<td>835 840 845</td>
</tr>
</tbody>
</table>
Arg Thr Phe His Lys Val Thr Leu Lys Asp Lys Met Ile Thr Val Thr
850     855     860
Arg Tyr Leu Pro Lys Tyr Pro Tyr Glu Ser Ala Gln Ile His Tyr Thr
865     870     875     880
Tyr Ser Leu Cys Pro Ser His Ser Asp Ser Glu Phe Val Ser Cys Trp
885     890     895
Val Glu Phe Ser His Glu Arg Leu Glu Glu Tyr Lys Trp Asn Tyr Leu
900     905     910
Asp Gln Tyr Ile Cys Ser Ala Gly Ser Glu Asp Phe Ser Leu Ile Glu
915     920     925
Ser Leu Lys Phe Trp Arg Thr Arg Phe Leu Leu Leu Pro Ala Cys Val
930     935     940
Thr Ala Thr Lys Arg Ile Thr Glu Gly Glu Ala His Cys Asp Ile Tyr
945     950     955     960
Gly Asp Arg Pro Arg Ala Asp Glu Asp Glu Trp Glu Leu Leu Asp Gly
965     970     975
Phe Val Arg Phe Val Glu Gly Leu Asn Arg Ile Arg Arg Arg Arg His Arg
980     985     990
Ser Asp Arg Met Met Arg Lys Gly Thr Ala Met Lys Gly Leu Gln Met
995    1000    1005
Thr Gly Pro Ile Ser Thr His Ser Leu Glu Ser Thr Ala Pro Pro
1010     1015     1020
Val Gly Lys Lys Gly Thr Ser Ala Leu Ser Ala Leu Leu Glu Met
1025     1030     1035
Glu Ala Ser Gln Lys Cys Leu Gly Glu Gln Gln Ala Ala Val His
1040     1045     1050
Gly Gly Lys Ser Ser Ala Gln Ser Ala Glu Ser Ser Ser Val Ala
1055     1060     1065
Met Thr Pro Thr Tyr Met Asp Ser Pro Arg Lys Val Ser Val Asp
1070     1075     1080
Gln Thr Ala Thr Pro Met Leu Asp Gly Thr Ser Leu Gly Ile Cys
1085     1090     1095
Thr Gly Gln Ser Met Asp Arg Gly Asn Ser Gln Thr Phe Gly Asn
1100    1105    1110
Ser Gln Asn Ile Gly Glu Gln Gly Tyr Ser Ser Thr Asn Ser Ser
1115    1120    1125
Asp Ser Ser Ser Gln Gln Leu Val Ala Ser Ser Leu Thr Ser Ser
1130    1135    1140
Ser Thr Leu Thr Glu Ile Leu Glu Ala Met Lys His Pro Ser Thr
1145 Gly Val Gln Leu Leu Ser Glu Gln Lys Gly Leu Ser Pro Tyr Cys
1150 1160 1165 1170
Phe Ile Ser Ala Glu Val Val His Trp Leu Val Asn His Val Glu
1170 1175 1180 1185
Gly Ile Gln Thr Gln Ala Met Ala Ile Asp Ile Met Gln Lys Met
1190 1195 1200
Leu Glu Glu Gln Leu Ile Thr His Ala Ser Gly Glu Ala Trp Arg
1200 1205 1210 1215
Thr Phe Ile Tyr Gly Phe Tyr Phe Tyr Lys Ile Val Thr Asp Lys
1220 1225 1230
Glu Pro Asp Arg Val Ala Met Gln Glu Pro Ala Thr Thr Trp His
1230 1235 1240 1245
Thr Ala Gly Val Asp Asp Phe Ala Ser Phe Glu Arg Lys Trp Phe
1250 1255 1260 1265
Glu Val Ala Phe Val Ala Glu Glu Leu Val His Ser Glu Ile Pro
1270 1275 1280 1285
 Ala Phe Leu Leu Pro Trp Leu Pro Ser Arg Pro Ala Ser Tyr Ala
1290 1295
Ser Arg His Ser Ser Phe Ser Arg Ser Phe Gly Gly Arg Ser Glu
1300 1305 1310 1315 1320 1325
 Ala Ala Ala Leu Leu Ala Ala Thr Val Pro Glu Glu Arg Thr Val
1330 1335
Thr Leu Asp Val Asp Val Asn Asn Arg Thr Asp Arg Leu Glu Trp
1340 1345 1350 1355 1360
Cys Ser Cys Tyr Tyr His Gly Asn Phe Ser Leu Asn Glu Ala Ala Phe
1370 1375 1380
Glu Ile Lys Leu His Trp Met Ala Val Thr Ala Ala Val Leu Phe
1385 1390 1395 1400 1405 1410
Glu Met Val Gln Gly Trp His Arg Lys Ala Thr Ser Cys Gly Phe
1410 1415 1420 1425 1430 1435 1440
Leu Tyr Gly Asp Pro Leu Arg Ala Gln Leu Phe Ile Pro Leu Asn
Leu Leu Val Pro Val Leu Glu Gly Pro Phe Ala Leu Pro Ser Tyr
1445 1450 1455 1460 1465 1470 1475 1480 1485 1490 1495 1500 1505
Ala Ile Ala His Arg Phe Gly Phe Val Gln Asp Lys Tyr Ser Ala
1445 1450 1455
Ser Ala Phe Asn Phe Pro Ala Glu Asn Lys Pro Gln Tyr Ile His
1460 1465 1470
Val Thr Gly Thr Val Phe Leu Gln Leu Pro Tyr Ser Lys Arg Lys
1475 1480 1485
Phe Ser Gly Gln Gln Arg Arg Arg Arg Ser Thr Ser Ser Thr
1490 1495 1500
Asn Gln Asn Met Phe Cys Glu Glu Arg Val Gly Tyr Asn Trp Ala
1505 1510 1515
Tyr Asn Thr Met Leu Thr Lys Thr Trp Arg Ser Ser Ala Thr Gly
1520 1525 1530
Asp Glu Lys Phe Ala Asp Arg Leu Leu Lys Asp Phe Thr Asp Phe
1535 1540 1545
Cys Ile Asn Arg Asp Asn Arg Leu Val Thr Phe Trp Thr Ser Cys
1550 1555 1560
Leu Glu Lys Met His Ala Ser Ala Pro
1565 1570

<210> 11
<211> 175
<212> PRT
<213> Homo sapiens

<400> 11

Met Leu Ser His Ser Ser Leu Thr Leu Ala Ala Pro Val Leu Cys Ala
1 5 10 15
Val Leu Ser Ser Leu Pro Trp Arg Trp Arg His Leu Cys Cys Val Pro
20 25 30
Cys Tyr Pro Thr Leu Leu Trp Arg Trp Arg His Leu Cys Cys Val Pro
35 40 45
Cys Tyr Pro Leu Phe Pro Gly Thr Gly Gly Thr Cys Ala Val Cys Arg
50 55 60
Val Thr Pro Leu Phe Pro Gly Ala Gly Gly Thr Cys Ala Met Cys Arg
65 70 75 80
Val Ile Leu Ser Ser Leu Ala Leu Val Ala Pro Val Leu Cys Ala Val
85 90
Leu Ser Ser Leu Pro Trp Arg Trp His Leu Cys Cys Val Leu Cys
100 105 110
Tyr Pro Leu Phe Pro Gly Ala Gly Gly Thr Cys Ala Met Cys Arg Val
115 120 125
Ile Leu Ser Ser Leu Ala Leu Ala Ala Arg Thr Leu Cys Ala Gly Val
130     135     140

Phe Thr Ser Ser Leu Trp Gly Ile Arg Leu Glu Thr Cys Phe Leu Pro
145  150  155  160

Ala Leu Lys Gly Cys Asn Ser Phe Val Leu Thr Val Pro Leu Asn
165  170  175