

CLAIMS

1. Method for predicting at least one fitness value of a protein, the method being implemented on a computer and including the following steps:

- encoding (100) the amino acid sequence of the protein into a numerical sequence according to a protein database (51), the numerical sequence comprising a value for each amino acid of the sequence;

- calculating (110) a protein spectrum from a Fourier Transform of the numerical sequence, the Fourier Transform being applied to the numerical sequence obtained further to the encoding step; and

for each fitness:

- comparing (130) the calculated protein spectrum with protein spectrum values of a predetermined database (55), said database containing protein spectrum values for different values of said fitness, wherein the comparison step (130) comprises determining, in the predetermined database (55) of protein spectrum values for different values of said fitness, the protein spectrum value which is the closest to the calculated protein spectrum according to a predetermined criterion, and

- predicting (130) a value of said fitness according to the comparison step, the predicted value of said fitness being then equal to the fitness value which is associated in said database with the determined protein spectrum value.

2. The method according to claim 1, wherein the calculated protein spectrum includes at least one frequency value and the calculated protein spectrum is compared with said protein spectrum values for each frequency value.

3. The method according to claim 1, wherein the Fourier Transform is a Fast Fourier Transform.

4. The method according to claim 1, wherein each protein spectrum verifies the following equation:

$$|f_j| = \left| \sum_{k=0}^{N-1} x_k \exp\left(\frac{-2i\pi}{N} jk\right) \right|$$

where j is an index-number of the protein spectrum $|f_j|$;

the numerical sequence includes N value(s) denoted x_k , with $0 \leq k \leq N-1$ and $N \geq 1$, k being an integer and N being a predetermined integer; and

i defining the imaginary number such that $i^2 = -1$.

5. The method according to any one of the preceding claims, wherein, during the encoding step (100), the protein database (51) includes at least one index of biochemical or physico-chemical property values, each property value being given for a respective amino acid; and

wherein, for each amino acid, the value in the numerical sequence is equal to the property value for said amino acid in a given index.

6. The method according to claim 5, wherein, during the encoding step (100), the protein database (51) includes several indexes of property values; and

wherein the method further includes a step of:

- selecting the best index based on a comparison of measured fitness values for sample proteins with predicted fitness values previously obtained for said sample proteins according to each index;

the encoding step (100) being then performed using the selected index.

7. The method according to claim 6, wherein, during the selection step, the selected index is the index with the smallest root mean square error,

wherein the root mean square error for each index verifies the following equation:

$$RMSE_{Index_j} = \sqrt{\sum_{i=1}^S \frac{(y_i - \hat{y}_{i,j})^2}{S}}$$

where y_i is the measured fitness of the i^{th} sample protein,

$\hat{y}_{i,j}$ is the predicted fitness of the i^{th} sample protein with the j^{th} index, and

S is the number of sample proteins.

8. The method according to claim 6, wherein, during the selection step, the selected index is the index with the coefficient of determination nearest to 1,

wherein the coefficient of determination for each index verifies the following equation:

$$R^2_{Index_j} = \frac{(\sum_{i=1}^S (y_i - \bar{y})(\hat{y}_{i,j} - \bar{\hat{y}}))^2}{\sum_{i=1}^S (y_i - \bar{y})^2 \sum_{i=1}^S (\hat{y}_{i,j} - \bar{\hat{y}})^2}$$

where y_i is the measured fitness of the i^{th} sample protein,

$\hat{y}_{i,j}$ is the predicted fitness of the i^{th} sample protein with the j^{th} index,

S is the number of sample proteins,

$\bar{y}$ is an average of the measured fitness for the S sample proteins, and

$\hat{\bar{y}}$ is an average of the predicted fitness for the S sample proteins.

9. The method according to any one of the preceding claims, wherein the method further includes, after the encoding step and before the protein spectrum calculation step, the following step:

- normalizing the numerical sequence obtained via the encoding step, by subtracting to each value of the numerical sequence a mean of the numerical sequence values;
the protein spectrum calculation step being then performed on the normalized numerical sequence.

10. The method according to any one of the preceding claims, wherein the method further includes, after the encoding step and before the protein spectrum calculation step, the following step:

- zero padding the numerical sequence obtained via the encoding step, by adding M zeros at one end of said numerical sequence, with M equal to $(N - P)$ where N is a predetermined integer and P is the number of values in said numerical sequence;
the protein spectrum calculation step being then performed on the numerical sequence obtained further to the zero padding step.

11. The method according to any one of the preceding claims, wherein, during the protein spectrum calculation step (110), several protein spectra are calculated for said protein according to several frequency ranges, and

wherein, during the prediction step, an intermediate value of the fitness is estimated for each protein spectrum according to the comparison step, and the predicted value of the fitness is then computed using the intermediate fitness values.

12. The method according to claim 11, wherein the predicted value of the fitness is computed with a regression on the intermediate fitness values.

13. The method according to claim 12, wherein the regression is a partial least square regression.

14. The method according to any one of the preceding claims, wherein the method includes a step of:

- analysis (140) of the protein according to the calculated protein spectrum, for screening of mutants libraries.

15. The method according to claim 14, wherein the analysis is done using a factorial discriminant analysis or a principal component analysis.

16. A computer program product including software instructions which, when implemented by a computer, implement a method according to any one of the preceding claims.

17. An electronic prediction system (20) for predicting at least one fitness value of a protein, the prediction system (20) including:

- an encoding module (50) configured for encoding the amino acid sequence into a numerical sequence according to a protein database (51), the numerical sequence comprising a value for each amino acid of the sequence;

- a calculation module (52) configured for calculating a protein spectrum from a Fourier Transform of the numerical sequence, the Fourier Transform being applied to the numerical sequence obtained by the encoding module; and

- a prediction module (56) configured for, for each fitness:

- + comparing the calculated protein spectrum with protein spectrum values of a predetermined database, said database containing protein spectrum values for different values of said fitness, wherein comparing comprises determining, in the predetermined database of protein spectrum values for different values of said fitness, the protein spectrum value which is the closest to the calculated protein spectrum according to a predetermined criterion, and

- + predicting a value of said fitness according to said comparison, the predicted value of said fitness being then equal to the fitness value which is associated in said database with the determined protein spectrum value.