Method for identifying proteins with N-terminal N-myristoylation

Background of the invention

Among the many known lipid modifications N-myristoylation is one of the best investigated and has been subject to several reviews (Towler et al. 1988b; Gordon et al. 1991; Han and Martinage 1992; Johnson et al. 1994; Boutin 1997). However, after the importance for cellular regulation, signal transduction, apoptosis (Zha et al. 2000) and the potential for diverse medical treatments (Felsted, Glover, and Hartman 1995; Parang et al. 1997; Sikorski et al. 1997; Gunaratne et al. 2000) have been rediscovered and the first structures of Myristoyl-CoA:Protein N-myristoyltransferases were published (Weston et al. 1998; Bhatnagar et al. 1998), this topic has gained increasing attention.

The rare C_{14} saturated fatty acid is linked most often cotranslationally (Olson and Spizz 1986; Wilcox, Hu, and Olson 1987) via an amid bond (Olson, Towler, and Glaser 1985) specific to the N-terminal glycine (Kamps, Buss, and Sefton 1985; Towler et al. 1987) of several eukaryotic and viral proteins. In general, the attachment of the lipid moiety results in an increase of hydrophobicity that plays an important role in membrane and protein association. Myristic acid represents less than 1% of all fatty acids in cells (Khandwala and Kasper 1971), but its specific length provides the

possibility for reversible interactions with other proteins or membranes (Peitzsch and McLaughlin 1993) in contrast to highly stable associations facilitated by other, more hydrophobic lipid modifications.

Myristoylation can be required but need not necessarily be sufficient for membrane anchoring, as known for example for the oncoprotein $p60^{v\text{-}src}$ (Resh 1994). The list of myristoylated proteins includes various kinases, phosphatases, cytochrome b_5 reductase, NO synthase, the α subunit of many G proteins, ADP ribosylation factors, the myristoylated alanine rich C kinase substrate (MARCKS) and other membrane- or cytoskeletal-bound structural proteins, Ca2+ binding/ EF hand proteins, as well as several viral proteins.

For the viruses, the lipid modification of the proteins is often essential by directing it to the plasma membrane of the host cell or it is necessary for the assembly of the viral structure (Moscufo, Simons, and Chow 1991) and replication in general (Raulin 2000). Many viruses take part in the pathological processes directly with their myristoylated oncoproteins (Arbelaez, Bernal, and Patarca 1999).

Myristoylation is not always reduced to a simple anchoring function. The fatty acid can also fold back to domains of the acylated protein itself and extend to the outside again controlled by the binding of Ca^{2+}-ions acting as a switch (Ames et al. 1996). Other examples of myristoyl switches for reversible membrane association (triggered by phosphorylation or proteolytic cleavage) can be found in MARCKS (McLaughlin and Aderem 1995) and the HIV-1 Gag precursor (Zhou and Resh 1996).

The most abundant form of myristoylation is catalyzed by myristoyl-CoA:protein N-myristoyltransferase (NMT) (Raju et al. 2000) which is absolutely specific to N-terminal glycine. The proposed invention focuses solely on the recognition of protein substrates for this enzymatic activity from the substrate's amino acid sequence.

The presently ongoing massive sequencing efforts result in an enormous amount of genomic data, which requires, in the next step, the detailed characterization of the encoded proteins. The above introduction to N-myristoylation emphasizes the immense importance of this lipid modification. Therefore in the post-genomic era, knowledge of a protein being myristoylated is invaluable additional information for its functional characterization.

The experimental procedures necessary for unambiguously identifying the lipid modification, often including the incorporation of ^{3}H-labeled myristic acid, are very laborious and time consuming. Additionally, the sequence similarity among related proteins is very high (often only single mutations within the region recognized by NMT) and, therefore, explicit experimental verification is often omitted.

Instead, the sequence is referred to a concordance with a consensus sequence of the myristoylation motif provided by available pattern search tools, e.g. PROSITE (Hofmann et al. 1999; Bucher and Bairoch 1994), which was last updated in April 1990. This pattern carries only a disproportionally small amount of the currently available information about the motif and produces a highly unrealistic number of

positive identifications of myristoylation sites and, with its current status, even false negative predictions.

The presently available single position based algorithms (such as the search with the PROSITE pattern) cannot deal with compensatory effects. This is an important reason why predictions of the myristoylation site have been insufficiently successful to date. The sequence context is much more critical than a unique disfavored residue. For example, while a single bulky amino acid at one position might well be tolerated in the size-limited groove where substrate recognition takes place, a second one in direct neighborhood probably will not. So, the closest positions will have to compensate the extra size of the predecessor.

Furthermore, the binding of a peptide substrate to an enzyme is a highly cooperative process in the sense of an induced fit. Therefore, not only consecutive amino acid residues but also positions further apart could have an influence on the binding of each other.

Extensive kinetic measurements of substrate specificity for the yeast NMT (Towler et al. 1988b) have emphasized this importance of the sequence context which could not be reflected by a simple pattern (such as the PROSITE motif).

NMT seems to be ubiquitous among eukaryotes. However, there exists an overlapping yet distinct (Towler et al. 1988a) substrate specificity between species, which became

obvious by several observations (Duronio et al. 1991). The PROSITE pattern contains no distinctive features to deal with the species-dependent substrate specificities.

Finally, using PROSITE to identify myristoylation sites is restricted to a binary decision (yes/no) without any estimation of the quality of the motif in the investigated sequence.

Due to the unsatisfactory performance of the currently available prediction tools for N-myristoylation, there is a need for an efficient and reliable tool that can significantly reduce the number of false positive predictions and that allows even large-scale database annotations. The proposed invention allows a reasonable pre-selection of candidate proteins and a dramatic speed up in sequence database searches aimed at the identification of proteins that have not previously been known to be myristoylated.

It was an objective of the present invention to provide a reliable method (prediction tool) to identify proteins that are candidates for N-terminal myristoylation, also from large protein sequence databases.

The solution of the problem underlying the present invention is based on the following considerations: By using approved statistical correlation methods, the characteristics for the positions in the N-terminus of all known substrates and, therefore, the obvious single position requirements for recognition by the enzyme are validated. Significance for compensatory effects are evaluated by fulfillment of the Fisher-criterion that characterizes correlation respectively independence of sequence positions.

To take advantage of the information collected above efficiently, a scoring function, validating the quality of a sequence as motif for recognition by NMT, should not be reduced to a term evaluating amino acid type preferences on single positions independently (e.g., as in profile approaches), but also comprise physical property restrictions as well as the compensatory effects from multiple sequence positions. Hence, a composite prediction function was created that combines profile-based sum scores using the PSIC algorithm (Sunyaev et al. 1999; S_{profile}) with special terms for the conserved physical properties, summarized as S_{ppt}.

$$S = S_{profile} + S_{ppt}$$

Implemented in a computer program, written in the C programming language, the user may choose the taxonomic parameter set and read in the sequence that should be investigated.

Then, S_{profile} is calculated for the query sequence and scored according to the profile extracted from a chosen learning set of known substrates. The conserved physical properties are assumed to follow a Gauss-like distribution when retrieved from the sequences already verified to be myristoylated. Deviation from the mean for the corresponding positions in the query sequence results in a penalty for the score proportional to the extent of the deviation. Thus, if nonconformity with the physical-chemical requirements for substrate recognition is more severe, the penalty will be higher.

The overall score is the sum of $S_{profile}$ and all the physical property terms (which are negative per definition as they are penalties). To address a threshold which sequences will be processed by NMT or not, the obtained scores are compared with the scores of the learning set and the lower limit for prediction is set to the lowermost score of an experimentally verified myristoylation site. The prediction is that proteins with higher scores should also be favored substrates of NMTs. At the same time, this simple approach cannot always reflect the naturally occurring different affinities to the enzyme in the relative size of prediction function scores.

For better judgment whether further experimental verification will be feasible for a candidate sequence and estimation of the general reliability of a prediction, the scores are translated into a probability of false-positive predictions applying generalized extreme value distribution functions (Eisenhaber et al. 2001).

In sequence space, the correct motif can occur incidentally with a certain probability. The more the scores of correctly predicted sequences differ from the scores of non-myristoylated proteins, the lower will be the probability for false-positive predictions and the higher will be the credibility of the prediction. The prediction method of the invention with its current parametrization learned from the today's available sequence examples allows even large-scale database annotations with less than 5 false positive assignments among 1000 unrelated sequences with an N-terminal glycine.

Summary of the Invention

In a first aspect, the present invention relates to a method for identifying candidate proteins with N-terminal N-myristoylation from the knowledge of their amino acid sequence, comprising the steps of

a) computational analysis of the N-terminal segment of the primary structure of the query protein and deciding the candidate's status by applying a scoring system based on

   (i) sensitive profile extraction,

   (ii) physical property requirements,

   (iii) compensatory effects over multiple positions, and

b) determining by experimental verification whether a protein selected in a) is a substrate for myristoyl-CoA:protein N-myristoyltransferase.

Step b) is conducted with proteins out of a reduced candidate list.

In a preferred embodiment, the scores obtained by step (a) are translated into probabilities of false-positive prediction (a1). Thus, by complementing the score function with rigorous statistics in form of a generalized extreme value distribution, a quality measurement of the myristoylation signal of investigated sequences can be provided.

The method of the invention can be carried out both in single target studies (e.g. studies aiming at identifying a protein as a therapeutic target) and in large-scale biomolecular sequence database scans.

In a second aspect, the present invention relates to a composite prediction algorithm combining profile-based sum scores ($S_{profile}$; (i)) with special terms for the conserved physical properties including compensatory effects, summarized as S_{ppt} ((ii) and (iii)).

$$S = S_{profile} + S_{ppt}$$

Sensitive profile extraction (i)

For a sensitive profile extraction (i), the PSIC algorithm (Sunyaev et al. 1999) is used, which is a powerful profile extraction technique that assigns both sequence- and alignment position-specific weights (Eisenhaber, Bork, and Eisenhaber 1999; Sunyaev et al. 1999). With this technique, the corrected relative occurrences $p(a,i)$ of amino acid types a at given motif positions i from the gapless multiple alignment of the N-termini (except the possible starting methionines) of the complete learning set of known myristoylated proteins is determined.

It should be emphasized that this multiple alignment contains many highly similar alignments with little sequence variations often affecting only a few positions. To

extract a maximum of information, an effective number $n(a,i)_{eff}$ of observations of amino acid type a at alignment position i is computed and finally determined $p(a,i)$ as

$$p(a,i) = \frac{n(a,i)_{eff}}{\sum_b n(b,i)_{eff}}.$$

The summation is carried out over all amino acid types b. The value $n(a,i)_{eff}$ is thought to depend on the overall similarity of sequences having the common amino acid type a in the alignment column considered. The frequency of identical alignment positions $f(a,i)$ in the subset of sequences having the same amino acid type a at alignment position i is used as similarity measure and is set equal to the probability of identical alignment positions for $n(a,i)_{eff}$ in random sequences. The solution of the equation

$$f(a,i) = \sum_b q_b^{n(a,i)_{eff}}$$

for $n(a,i)_{eff}$ estimates the number of independent observations of amino acid a at position i in the alignment. The value q_b is the default frequency of amino acid type b in a sequence database.

The final profile matrix $S_i(a)$ is calculated as (Sunyaev et al. 1999)

$$S_i(a) = \ln \frac{p(a,i)}{q_a}.$$

These values represent the subscores of single positions i that sum up to regional scores within their defined regions.

$$S_{region} = \sum_{i \in region} S_i$$

The regional subscores finally enter $S_{profile}$ adjusted with a weighting factor α_{region}, emphasizing the importance of key positions, and a normalization condition $\alpha_{profile}$, compensating for the different lengths of the sequence regions.

$$\alpha_{profile}^{-1} \cdot profile_length = \sum_{region} \alpha_{region} \cdot region_length$$

The complete profile score term $S_{profile}$ can therefore be circumstantiated as follows:

$$S_{profile} = \sum_{regions} \alpha_{profile} \alpha_{region} S_{region}$$

The profile is computed over about 40 amino acids following the obligatory N-terminal glycine for all entries of the learning set. The scores for query sequences are calculated only for positions 2 to 17. The latter position range corresponds to sequence segments showing detectable property deviations in N-terminally N-myristoylated proteins compared with unrelated sequences.

Physical property requirements and compensatory effects ((ii) and (iii))

The functional form of multiple residue correlation terms with respect to physical properties (P) composing S_{ppt} is selected in such a manner that clear deviations from value ranges in the learning set are penalized. At the same time, compliance with the

consensus signal extracted from the learning set results in a zero score (but not in positive scores).

$$T_j(P) = \begin{cases} 0 & if \quad P \le P_j \\ -(P - P_j)^2 / (2\sigma_j^2) & if \quad P > P_j \end{cases}$$

Herewith, it is recognized that the form of physical terms in S_{ppt} may reflect the present rough understanding of requirements of the polypeptide binding site in myristoyl-CoA:protein N-myristoyltransferase. A possibly specific role of different amino acid types at certain sequence positions might be not well discerned. Definitely, it can differ among species. In its current formulation, S_{ppt}

$$S_{ppt} = \sum_{j=0}^{11} \alpha_j T_j$$

includes the following terms describing

- Side-chain volume limitations and mutual volume compensations for residues on positions 2 and 3, as well as compensation of bulkiness between position 7 and 9, respectively overall size limitation (positions 2 to 11) within the catalytic cleft (terms T_0, T_3 and T_8);

- Limited extent of hydrophobicity on positions 2 and 3, hydrophobicity compensations on positions 8, 9 and 10, as well as between positions 2 and 5 (terms T_1, T_4 and T_7);

- Spacer region (position 6 to 17) that should not contain too many hydrophobic residues, evaluated by the average of optimal matching hydrophobicity (term T_2);
- Flexibility compensations in the selective binding region of positions 3 to 5 (term T_5);
- Certain polarity requirements for positions 5 and 6 (term T_6).

Each of these nine physical conditions is weighted according to its importance and enters the sum above in form of the natural logarithm of a probability distribution function to be comparable with scores from profile computations. A Gauss-like distribution was assumed for values outside the allowed value ranges (see formula for $T_j(P)$ above!).

Furthermore, there are penalties with fixed thresholds that are also added to S_{ppt}:

- To exclude signalpeptide sequences, the typical structural features of their hydrophobic region are penalized when the hydrophobicity in a sequence window of 4 amino acids exceeds a certain threshold (term T_9 and T_{10});
- Penalties are also given for extraordinary residues on special positions in concordance with the compositional analysis and the kinetic measurements (term T_{11}).

The purpose of introducing S_{ppt} consists in excluding sequences as unlikely candidates for myristoylation due to untypical integral sequence properties compared with the learning set.

Estimation of the probability of false positive prediction (a1)

Furthermore, it was sought to facilitate the interpretation of the scores generated by the program of step a) for users that are not familiar with the prediction function and to make the obtained scores comparable with scores calculated for other features (for example prediction of GPI-anchor sites; Eisenhaber et al., 1999) This problem was solved by further applying, in a preferred embodiment of the invention, step (a1) in the framework of statistical theory by calculating the probability of false positive predictions, i.e., the probability of incidental occurrences of a motif match with the same or better score in an unrelated sequence. If a score S is normally distributed, then the probability P of a score S to be larger than a threshold S_{th} is described by an extreme-value distribution (Altschul et al. 1994) with generalized analytical form (Eisenhaber et al. 2001):

$$P(S > S_{th}) = 1 - e^{-e^{-f(S_{th})}}$$

$$f(S_{th}) = \sum_{i=1}^{n} \lambda_i (S_{th} - u)^i$$

The scores for sets of unrelated sequences (non-myristoylated proteins) for the respective taxonomic groups are calculated and the best reasonable polynomial fit (i is between 2 and 6) is evaluated to describe the generalized extreme value distribution.

So the score of the investigated sequence can be converted into a probability of false positive prediction.

The method of the invention distinguishes three groups of sequences: The first one consists of the sequences that obtain positive scores. Their probability of being false positive is maximum 0.5% for the different taxonomic parameter sets. Based on the present understanding of the requirements of the NMT binding pocket, they may be predicted as a "certain" myristoylation site. The second group can be seen as a twilight zone, because there are also verified myristoylation sites with scores between –2 and 0. They can be predicted as "probable", as the probability of having a score > –2 is accordingly higher (2,8% maximum). Sequences with scores lower than –2 represent the third group and will not be predicted by the program of the invention as protein candidates for N-myristoylation.

Experimental verification of N-myristoylation

Whether a selected protein is a substrate for myristoyl-CoA:protein N-myristoyltransferase, can be experimentally confirmed according to step b) by methods known *per se* in the art, in particular, if the number of candidates is not too large.

Several methods that are useful in the present invention to verify the attachment of myristic acid to a protein have been described in the literature. For the purpose of identifying large numbers of myristoylated proteins from different tissues and all

possible organisms, an in vitro bacterial co-expression system (Duronio et al. 1990; Knoll et al. 1995) to obtain both NMT and the candidate protein in recombinant form (sequentially expressed in *Escherichia coli*) can be used (see Example 6). Although this method only indicates in vitro myristoylation of a candidate protein, it clearly states whether the protein can be a substrate for NMT.

Isolation and purification strongly depend on the characteristics of the protein (e.g. lipophilicity, subcellular compartment: membrane-bound, cytosolic...). Therefore, it is important to follow procedures that have been reported to be suitable for the properties of the protein of interest.

In order to determine the myristoylation of viral proteins, the method of (Schultz and Oroszlan 1983) (see Example 6 below) can be used.

A frequently employed protein purification method is SDS-PAGE, followed by further identification by immunoprecipitation with a specific antibody against the candidate protein. In this method, incorporation of ^{3}H-labeled myristate (via addition to the medium) helps to monitor the proteins of interest during isolation and purification processes.

A more accurate method for identifying N-myristoylation is the use of FAB-MS (Fast Atom Bombardment Mass Spectrometry; (Carr et al. 1982) or combined analytical methods (Neubert and Johnson 1995), such as GC-MS (Gas Chromatography - Mass Spectrometry) or HPLC-ESI-MS (High Pressure Liquid Chromatography –

Electrospray Ionization – Mass Spectrometry). In this method, N-terminal N-myristoylation is ascertained by comparing the molecular weight obtained by mass spectrometry of a protein or its fragments with calculated masses based on the amino acid sequence of the protein. The discrepancy in weights has to correspond to the mass of the myristoyl anchor.

Other appropriate examples for methods that may be employed in the present invention are the expression and characterization of calcium-myristoyl switch proteins (Zozulya, Ladant, and Stryer 1995) or ADP-ribosylation factors (Randazzo and Kahn 1995). For high-resolution structural determination of protein-linked acyl groups it is referred to (Neubert and Johnson 1995). More detail is contained in example 6.

Validation of the prediction method

The inventors executed several acknowledged tests and comparisons between predicted and experimental data to validate the methodology proposed in this invention which are listed below. Each test is explained in more detail in the following section and/or in one of the examples at the end of this document.

- Self-Consistency test (example 1)
- Jack-knife test over whole score $\boldsymbol{S}$ (example 2)
- Jack-knife test over $\boldsymbol{S}_{ppt}$, while $\boldsymbol{S}_{profile}$ was calculated with the whole learning set (example 2)

- Scores for proteins that are reported not to be myristoylated (example 3)
- Correlation with experimental kinetic data (example 4)
- Comparison with alternative prediction methods (example 5)

In order to clarify whether the chosen parameter sets describe the available information on N-terminal myristoylation sufficiently, it was investigated whether the chosen learning set can be predicted itself (self-consistency test).

Many prediction techniques suffer from over-parameterization; i.e., a too small learning set may be over-fitted with a too large set of parameters so that it perfectly predicts the present learning data but actually fails to describe the biological problem in its entirety. As a result, only sequences closely sequentially related to the learning set would be predicted. The so-called jack-knife or cross-validation test is designed to test for such cases. In this scheme, the predicted sequence is excluded from the learning procedure.

Two types of jack-knife tests have been applied. In the first test, the predicted sequence was excluded from the complete learning procedure (including profile derivation). In a second variant of the jack-knife test, the profile matrix calculation was executed with all sequences but the predicted sequence was left out for the calculation of S_{ppt} only. Whereas the first jack-knife test checks the whole procedure for parameter over-fitting, the second test is a specific control for the parameters of the chosen physical property terms.

The scientific literature provides lists of proteins that are reported not to be myristoylated despite of their N-terminal glycine. The performance of the method of the invention was also tested for these potential candidates for false positive predictions.

Finally, the regression between affinity constants for several octapeptides towards the enzyme and the respective scores generated by the program was analyzed.

In summary, the present invention provides a novel method for the identification of N-terminal N-myristoylation that only relies on the primary structure of proteins. This technique is also useful for the selection of targets from large biomolecular sequence databases.

The central element of the invention, the novel algorithm implemented in a software tool, not only evaluates amino acid type preferences on single positions with the powerful profile extraction method PSIC, but also penalizes deviations from the physical-chemical requirements for several positions that either show a general trend of deviation from the average properties of known myristoylated proteins or have been tested for compensatory effects including multiple sequence positions. For the quantification of the risk of false positive assignment, the scores were translated into probabilities of false-positive prediction using a rigorous statistical approach. Given the high reliability of the method proposed in this invention, it becomes feasible for the first time to scan large biomolecular sequence databases and to annotate protein substrates for N-terminal myristoylation automatically. Thus, the proposed invention is

a precious tool in the post-genomic era where the enormous amount of data can only be mastered with computational approaches.

Example 1

Self-consistency test

The information of the myristoylation motif of a set of test candidate proteins has been derived from a selected learning set. However, to check whether this knowledge has been converted successfully into the parameters of the prediction method of the invention, it was examined how the program would predict the learning set itself.

The program of the invention was shown to be capable to predict 359 of the 368 (97.6%) entries within the Eukaryota minus Fungi plus Viruses set. With the fungal specific parameters, which are very restrictive to maintain consistency with different substrate specificity, all 22 fungal entries (100%) were positively predicted.

The 9 entries that could not be predicted (see table 1) attract attention because they show major discrepancies with the consensus description of the motif. Several ARFs (ADP ribosylation factors) have been shown to be myristoylated (Randazzo and Kahn 1995). However, the entries listed in Table 1 do not share the motif of the experimentally verified proteins and are annotated only as potential (score lower than -2).

Penalties by the program arise because of the three consecutive large hydrophobic residues that should be hard to accommodate in the size limited pocket harboring

substrate positions 2, 3 and 4. Isoleucine or valine on the critical position 5 which is situated within a narrow ring of negatively charged aspartates should also be strongly disfavored. Mutation of octapeptide GNAAAARR to GPAAAARR resulted in loss of the ability to become myristoylated (Towler et al. 1988b). Therefore, proline on 2 in STK_HYDAT (P17713) can have a similar effect by dramatically reducing the flexibility of the substrate.

The obtained results show that these 9 entries do not match the N-myristoylation consensus pattern. Therefore, it is assumed that these proteins cannot be good substrates for NMT.

Example 2

Jack-knife tests

In the first approach, the sequence that should be predicted was left out of the complete learning procedure (both profile and physical property calculations). The method used for profile extraction with sequence and position-specific weightings has the advantage that in spite of the subsets with similar sequences even small differences contribute to the extracted information. Therefore, the values calculated for the profile matrix vary even when a sequence from a subset of homologues is left out. For the large set of eukaryotes and viruses, but without fungi, 353 out of 368 (95.9%) sequences are still predicted to have a myristoylation site. From the smaller fungal set 21 of 22 (95.5%)

sequences remain recognized by the program. As expected, the entries that failed to be predicted within the self-consistency test are once again out of the prediction limit. In addition, some very distinct but still trustworthy sequences produce a major shift in the profile matrix when they are left out and cannot be predicted anymore without their contribution to the profile, as expected.

The second jack-knife test shows that S_{ppt} parameters are stably derived from the learning set. Therefore, the sequence that should be predicted was left out of the calculation of S_{ppt} but not $S_{profile}$, resulting in a constant profile matrix throughout the whole test. The fact that for both parameter sets the prediction accuracies from the self-consistency test were reached (97.6% for EUKARYOTA-FUNGI+VIRUSES, respectively 100.0% for FUNGI) signifies that the learning set is large enough for a reliable determination of the physical property terms. At the same time, the results of the first jack-knife test show that the learning set of sequences is still too small for a stable derivation of the profile parameters. An enlarged learning set that may become available in the future will further improve the rate of correct positive prediction. It should be noted that, without making any changes to the approach outlined in this description and without modification of the associated software, an updated learning set including also protein sequences with experimentally verified N-terminal myristoylation available at a later date can be applied with the method presented.

Example 3

Scores for proteins that have been reported not to be myristoylated

One of the reviews (Boutin 1997) lists a table based on data by Tsunasawa (Tsunasawa, Stewart, and Sherman 1985) with proteins that have been investigated for myristoylation, but did not have myristic acid as attachment despite of their N-terminal glycines. Table 2 presents the scores that the method of the invention generates for these sequences. In complete concordance with the experiments, none of them would be positively predicted. (Table 2 is based on data by Tsunasawa (Tsunasawa, Stewart, and Sherman 1985) and Boutin (Boutin 1997)). It lists proteins that were reported not to be myristoylated with the scores they obtain by the algorithm of the invention. The threshold for prediction with the attribute 'probable' is a score higher than –2.0 and, thus, none of them does have a myristoylation site according to the present understanding of the motif and in agreement with the literature.

Example 4

Correlation of predicted N-terminal N-Myristoylation with experimental kinetic data

The enzyme Myristoyl-CoA:protein N-myristoyltransferase from Saccharomyces cerevisiae has been subject to extensive kinetic measurements to examine its substrate specificity (Towler et al. 1988b). Reliably quantifying the affinity constants or

myristoylation velocities is a difficult task as can be seen by crude discrepancies within the provided data. For example, the K_m of octapeptides GNAAAARR and GSSKSKPK were specified with 31, respectively 105 μM in one paper (Rocque et al. 1993), while an earlier publication lists other values for the same octapeptides ($K_{m\,[GNAAAARR]}$ = 60 μM; $K_{m\,[GSSKSKPK]}$ = 40 μM) (Towler et al. 1988b). It is clear that these results can only be compared within one series of experiments as they are depending on many factors (like temperature, solutions...) and experiments by different people, although from the same group, will not provide exactly the same values. However, the relation of the affinities, certifying which substrates are binding better than others, should remain constant.

In the given case, the ratio of the K_m even topples, attesting GNAAAARR a 3-fold higher affinity (lower K_m) than GSSKSKPK, controversial to previous data where GSSKSKPK was supposed to bind better than GNAAAARR.

Nevertheless, the scores predicted by the method of the invention coincide at least qualitatively with most of the experimental data (Figure: Correlation between the negative logarithm of experimentally derived K_m from yeast substrates and the scores calculated with the fungal parameter set). Following the more recent publication (Rocque et al. 1993) that was also used for discussion of the species dependent substrate specificities, the scores for the respective octapeptides were calculated, which all had to be extended with a common linker region. Single mutations in the derivates of cAMP-dependent protein kinase GNAAAARR cause an increase or decrease in the

score proportional to the negative logarithm of the affinity. But also GARASVLS maintains the proportion, showing that not only single position deviations, but also the sequence context (after multiple mutations) is recognized at least in part by the algorithm.

Unfortunately, GLYASKLS as well as GSSKSKPK suffer from the profile parameterization because of their infrequent amino acids at position 2 to 4. Of course they are predicted positively, but their scores do not fulfill the order of affinity derived from the experiment in context of the other sequences. Finally, the score for GSSKSKPK would suit better for the values published earlier, raising the question which experimental results are more correct.

The remaining sequences however, show the expected behavior with a higher score for substrates with higher affinity to the enzyme (Figure).

Example 5

Comparison of the novel prediction function with commonly used prediction methods

So far, the only available tool for prediction of myristoylation sites has been the pattern search with PROSITE. Apart from its value for several other applications, it is not adequate given the complexity of this special motif. Last updated over 10 years ago, the concept of allowing only a limited subset of amino acid types on single positions will never be able to reflect the dependence on the sequence context for the recognition by NMT.

In its current status, this approach even produces false negative predictions for proteins whose myristoylation has been verified by experiment (ARF6_HUMAN [P26438], ARF6_CHICK [P26990], GCAP_BOVIN [P46065], HIA1_DICDI [P13231], HIA2_DICDI [P42526], HIPP_HUMAN [P41211], HIPP_RAT [P32076], NCAH_DROME [P42325], NECD_BOVIN [P29554], NECX_APLCA [Q16982] and many others that share broad similarity with these sequences...). Of course, updating of the PROSITE motif will correct this failure, but as a consequence also dramatically boost the number of false positives, because the allowance of further amino acids makes the weak pattern even more common.

In comparison to the predictions for proteins reported not to be myristoylated (table 2, 100% correctly not predicted), the PROSITE pattern recognition would predict myristoylation sites for 3 of these sequences (62.5% correct), respectively even 5 (only 37.5% correct) after a conscientious update.

Table 3 shows the performance of the tool of this invention compared to a PROSITE pattern search with the current motif (old, producing false negative predictions) and a corrected version (updated) over a large database. The same table lists also the number of predictions of a PROSITE pattern search over the complete SWISSALL database compared to our algorithm. For fair comparison, the PROSITE search was restricted to N-terminal glycines (possibly after a leading methionine), which is not implemented in the usual PROSITE pattern, thereby producing the highly unrealistic numbers of predictions. As PROSITE cannot distinguish between any taxon-specific substrate

preferences, the prediction algorithm was reduced to the less stringent parameter set derived from a learning set containing eukaryan and viral sequences but none from fungi.

Finally, by using the information of already known myristoylated proteins restricted to certain amino acids, the PROSITE approach will never be capable of finding distinct motifs, where mutation of residues still leads to viable substrates for NMT due to compensatory effects. As can be seen from the comparison. The problem of sequence positional correlation cannot be addressed with a simple profile or pattern search alone.

Example 6

Experimental verification of the predicted myristoylation

The computational approach has the practical advantage that, among the many available proteins, examples unrelated to N-terminal myristoylation can be unselected. Thus, the resulting list of candidates becomes very small. The high-scoring hits are certain substrates for this protein modification. It is also feasible to check the remaining hits with lower scores for N-myristoylation experimentally with existing techniques.

There are several ways to verify the attachment of myristic acid to a protein. For the purpose of identifying large numbers of myristoylated proteins from different tissues

and all possible organisms an in vitro bacterial coexpression system using recombinant NMT and candidate protein (sequentially expressed in *Escherichia coli*) is very efficient (Duronio et al. 1990). However, this only indicates in vitro myristoylation of a candidate protein, but nevertheless clearly states whether it can be a substrate to NMT. For details of this elaborated method see (Knoll et al. 1995).

1. Determining N-myristoylation of eukaryotic proteins by using an E. coli coexpression system (Knoll et al., 1995)

a) *E. coli* is transformed with plasmids that direct expression of NMT (Weston et al. 1998; Bhatnagar et al. 1998) and candidate protein. Next, simultaneous resistance to ampicillin and kanamycin is selected for (100 μg/ml of each in Luria broth plates).

b) Plasmid DNA from transformants is prepared and restriction endonuclease digestions used to verify that both plasmids are present.

c) A 20-ml culture of double transformant is grown at 37 °C in Luria broth plus antibiotics to A_{600} of 0.5-0.7 and IPTG is added to 1 m*M* (NMT induction).

d) The culture is incubated for 40 min, then nalidixic acid is added at 50 μg/ml (candidate protein induction).

e) 2-ml aliquots of culture are immediately added to polypropylene tubes containing [^{3}H]myristate (50-100 μCi/ml culture; label should be dried in tube under N_2 prior to addition of culture).

f) The cells are incubated for 30 min at 37 °C, then placed on ice for 5 min.

g) The cells are collected by centrifugation at 4000 g for 10 min and washed with 1 ml of phosphate-buffered saline (PBS). The cell suspension is transferred to a microcentrifuge tube and centrifuged once again.

h) The cells are resuspended in 50 μl lysis buffer (0.24 *M* Tris, pH 6.8, 2% SDS/ml culture), boiled for 5 min and then centrifuged for 5 min. The supernatant is saved; total protein content is determined using BCA protein assay (Pierce, Rockford, IL). Final analysis of myristyolation is done by SDS-PAGE (load 100 μg protein/lane) and fluorography.

2. Determining N-myristoylation of viral proteins (Schultz and Oroszlan 1983):

a) Radiolabeling of cell cultures

[^{3}H]-myristic acid (0.5 mCi) is dissolved in a minimum volume of 70% ethanol and diluted to 1 ml with Eagle modified essential medium supplemented with 10% fetal calf serum. A T25 flask of each virus-producing cell line is labeled with 1 ml of this medium for 5 min. At the end of the labeling period, cells are rinsed, lysed, and immunoprecipitated with the appropriate monospecific antiserum and protein A-Sepharose (Pharmacia Fine Chemicals, Piscataway, N.J.) as described (Schultz,

Rabin, and Oroszlan 1979). Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) with linear-gradient 7.5 to 17% polyacrylamide gels follows the formulation of (Laemmli 1970). Gels are impregnated with PPO (2,5-diphenyloxazole; Amersham Corp., Arlington Heights, III.) for fluorography (Laskey and Mills 1975).

b) Purification of viral proteins:

Viral proteins are labeled and purified by reverse-phase high-performance liquid chromatography as follows. Virus particles are recovered from the medium of cultures which have been labeled overnight with [^{3}H]-myristate (0.5 mCi/ml) by pelleting through a cushion of 20% sucrose in TNE buffer (0.05 M Tris-hydrochloride [pH 7.5], 150 mM NaCl, 1 mM EDTA) for 90 min at 105000·*g*. The drained pellet is dissolved in 6 M guanidine hydrochloride, adjusted to pH 2 with trifluoroacetic acid, and applied to a μ Bondapak C_{18} column (Waters Associates, Milford, Mass.). Gradient elution chromatography is accomplished with a Waters Associates model 660 solvent programmer, two model 6000M solvent delivery pumps, and a model 450 variable wavelength detector set at 206 nm. Solvents for reverse-phase high-performance liquid chromatography are 0.1% trifluoroacetic acid or propanol with 0.1% trifluoroacetic acid as the organic phase. Proteins are eluted in a 1-h linear gradient from 0 to 60% acetonitrile at ambient temperature followed by a 15-min linear gradient from 20 to 60% propanol at 50°C (Henderson, Sowder, and Oroszlan 1981).

Annotated entries not predicted with the function		
SWISSPROT ID	**N-terminal sequence**	**SWISSPROT-Annotation**
ARF1_BRARP (Q96361)	GILFTRMFSSVFGNKEARILVLGLDNAGKTTILYRLQMGEV	Potential
ARF1_PLAFO (Q25761)	GLYVSRLFNRLFQKKDVRILMVGLDAAGKTTILYKVKLGEV	Potential
ARF3_ARATH (P40940)	GILFTRMFSSVFGNKEARILVLGLDNAGKTTILYRLQMGEV	Potential
ARF3_CAEEL (Q94231)	GLFFSKISSFMFPNIECRTLMLGLDGAGKTTILYKLKLNET	Potential
ARFL_CAEEL (P34212)	GLIMAKLFQSWWIGKKYKIIVVGLDNAGKTTILYNYVTKDQ	Potential
ARF_PLAFA (Q94650)	GLYVSRLFNRLFQKKDVRILMVGLDAAGKTTILYKVKLGEV	Potential
ENTK_PIG (P98074)	GSKRIIPSRHRSLSTYEVMFTALFAILMVLCAGLIAVSWLT	Potential
NB8M_HUMAN (P17568)	GAHLVRRYLGDASVEPDPLQMPTFPPDYGFPERKEREMVAT	by similarity
STK_HYDAT (P17713)	GPCCSKQTKALNNQPDKSKSKDVVLKENTSPFSQNTNNIMH	by similarity

Table 1

Proteins reported not to be myristoylated				
ID	**N-terminal Sequence**	**Score**	**Probability of FALSE positive prediction (%)**	**Reference**
OVAL_CHICK (P01012)	GSIGAASMEFCFDVFKELKV	-3.178	5.61	(Palmiter, Gagnon, and Walsh 1978)
LGB2_SOYBN (P02236)	GAFTEKQEALVSSSFEAFKA	-4.546	10.44	(Hyldig-Nielsen et al. 1982)
HBG1_PONPY (P18995)	GHFTEEDKATITSLWGKVNV	-17.735	67.46	(Slightom, Blechl, and Smithies 1980)
CRG2_SCOWA (P19865)	GKITFYEDRNFQGRCYECST	-7.139	22.48	(Chiou et al. 1990)
MYG_HALGR (P02162)	GLSDGEWHLVLNVWGKVETD	-22.332	77.84	(Blanchetot et al. 1983)
CYC_APTPA (P00017)	GDIEKGKKIFVQKCSQCHTV	-17.723	67.43	(Limbach and Wu 1983)
GBAS_BOVIN (P04896)	GCLGNSKTEDQRNEEKAQRE	-5.809	16.02	(McCallum et al. 1995)
ACT1_ACACA (P02578)	GDEVQALVIDNGSGMCKAGF	-15.954	61.44	(Vandekerckhove, Lal, and Korn 1984)

Table 2

Number of predictions with PROSITE pattern search and with our algorithm for the complete SWISSALL database			
90939 sequences (19.01.2001)		(M)G	all G (unrealistic)
PROSITE	old	1172	450986
	updated	1726	685963
Our algorithm		699	

Table 3

References

1. Altschul, S. F. et al. 1994 "Issues in searching molecular sequence databases." *Nat.Genet.* 6(2): 119-129.

2. Ames, J. B. et al. 1996 "Portrait of a myristoyl switch protein." *Curr.Opin.Struct.Biol.* 6(4): 432-438.

3. Arbelaez, A. M., C. Bernal, and R. Patarca. 1999 "Acute retroviruses and oncogenesis." *Crit Rev.Oncog.* 10(1-2): 17-81.

4. Bhatnagar, R. S. et al. 1998 "Structure of N-myristoyltransferase with bound myristoylCoA and peptide substrate analogs." *Nat.Struct.Biol.* 5(12): 1091-1097.

5. Bhatnagar, R. S. et al. 1999 "The structure of myristoyl-CoA:protein N-myristoyltransferase." *Biochim.Biophys.Acta* 1441(2-3): 162-172.

6. Blanchetot, A. et al. 1983 "The seal myoglobin gene: an unusually long globin gene." *Nature* 301(5902): 732-734.

7. Boutin, J. A. 1997 "Myristoylation." *Cell Signal.* 9(1): 15-35.

8. Bucher, P. and A. Bairoch. 1994 "A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation." *Ismb.* 2: 53-61.

9. Carr, S. A. et al. 1982 "n-Tetradecanoyl is the NH2-terminal blocking group of the catalytic subunit of cyclic AMP-dependent protein kinase from bovine cardiac muscle." *Proc.Natl.Acad.Sci.U.S.A* 79(20): 6128-6131.

10. Chiou, S. H. et al. 1990 "Comparison of the gamma-crystallins isolated from eye lenses of shark and carp. Unique secondary and tertiary structure of shark gamma-crystallin." *FEBS Lett.* 275(1-2): 111-113.

11. Duronio, R. J. et al. 1990 "Protein N-myristoylation in Escherichia coli: reconstitution of a eukaryotic protein modification in bacteria." *Proc.Natl.Acad.Sci.U.S.A* 87(4): 1506-1510.

12. Duronio, R. J. et al. 1991 "Analyzing the substrate specificity of Saccharomyces cerevisiae myristoyl-CoA:protein N-myristoyltransferase by co-expressing it with

mammalian G protein alpha subunits in Escherichia coli." *J.Biol.Chem.* 266(16): 10498-10504.

13. Eisenhaber, B., P. Bork, and F. Eisenhaber. 1999 "Prediction of potential GPI-modification sites in proprotein sequences." *J.Mol.Biol.* 292(3): 741-758.

14. Eisenhaber B, Bork P, Eisenhaber F. 2001 "Post-translational GPI lipid anchor modification of proteins in kingdoms of life: analysis of protein sequence data from complete genomes." *Protein Eng.* 14(1):17-25.

15. Felsted, R. L., C. J. Glover, and K. Hartman. 1995 "Protein N-myristoylation as a chemotherapeutic target for cancer [editorial]." *J.Natl.Cancer Inst.* 87(21): 1571-1573.

16. Gordon, J. I. et al. 1991 "Protein N-myristoylation." *J.Biol.Chem.* 266(14): 8647-8650.

17. Guex, N. and M. C. Peitsch. 1997 "SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling." *Electrophoresis* 18(15): 2714-2723.

18. Gunaratne, R. S. et al. 2000 "Characterization of N-myristoyltransferase from Plasmodium falciparum." *Biochem.J.* 348 Pt 2: 459-463.

19. Han, K. K. and A. Martinage. 1992 "Post-translational chemical modification(s) of proteins." *Int.J.Biochem.* 24(1): 19-28.

20. Henderson, L. E., R. Sowder, and S. Oroszlan. 1981. "Protein and peptide purification by reverse-phase high-pressuer chromatography using volatile solvents." Pp. 251-260 in *Chemical synthesis and sequencing of peptides and proteins.* Edited by D. T. Liu, A. N Schecter, and R. Condliffe. Amsterdam: Elsevier Nederland.

21. Hofmann, K. et al. 1999 "The PROSITE database, its status in 1999." *Nucleic Acids Res.* 27(1): 215-219.

22. Hyldig-Nielsen, J. J. et al. 1982 "The primary structures of two leghemoglobin genes from soybean." *Nucleic Acids Res.* 10(2): 689-701.

23. Johnson, D. R. et al. 1994 "Genetic and biochemical studies of protein N-myristoylation." *Annu.Rev.Biochem.* 63: 869-914.

24. Kamps, M. P., J. E. Buss, and B. M. Sefton. 1985 "Mutation of NH2-terminal glycine of p60src prevents both myristoylation and morphological transformation." *Proc.Natl.Acad.Sci.U.S.A* 82(14): 4625-4628.

25. Khandwala, A. S. and C. B. Kasper. 1971 "The fatty acid composition of individual phospholipids from rat liver nuclear membrane and nuclei." *J.Biol.Chem.* 246(20): 6242-6246.

26. Knoll, L. J. et al. 1995 "Functional significance of myristoyl moiety in N-myristoyl proteins." *Methods Enzymol.* 250: 405-435.

27. Laemmli, U. K. 1970 "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." *Nature* 227(259): 680-685.

28. Laskey, R. A. and A. D. Mills. 1975 "Quantitative film detection of 3H and 14C in polyacrylamide gels by fluorography." *Eur.J.Biochem.* 56(2): 335-341.

29. Limbach, K. J. and R. Wu. 1983 "Isolation and characterization of two alleles of the chicken cytochrome c gene." *Nucleic Acids Res.* 11(24): 8931-8950.

30. McCallum, J. F. et al. 1995 "The role of palmitoylation of the guanine nucleotide binding protein G11 alpha in defining interaction with the plasma membrane." *Biochem.J.* 310 (Pt 3): 1021-1027.

31. McLaughlin, S. and A. Aderem. 1995 "The myristoyl-electrostatic switch: a modulator of reversible protein- membrane interactions." *Trends Biochem.Sci.* 20(7): 272-276.

32. Moscufo, N., J. Simons, and M. Chow. 1991 "Myristoylation is important at multiple stages in poliovirus assembly." *J.Virol.* 65(5): 2372-2380.

33. Neubert, T. A. and R. S. Johnson. 1995 "High-resolution structural determination of protein-linked acyl groups." *Methods Enzymol.* 250: 487-494.

34. Olson, E. N. and G. Spizz. 1986 "Fatty acylation of cellular proteins. Temporal and subcellular differences between palmitate and myristate acylation." *J.Biol.Chem.* 261(5): 2458-2466.

35. Olson, E. N., D. A. Towler, and L. Glaser. 1985 "Specificity of fatty acid acylation of cellular proteins." *J.Biol.Chem.* 260(6): 3784-3790.

36. Palmiter, R. D., J. Gagnon, and K. A. Walsh. 1978 "Ovalbumin: a secreted protein without a transient hydrophobic leader sequence." *Proc.Natl.Acad.Sci.U.S.A* 75(1): 94-98.

37. Parang, K. et al. 1997 "In vitro antiviral activities of myristic acid analogs against human immunodeficiency and hepatitis B viruses." *Antiviral Res.* 34(3): 75-90.

38. Peitzsch, R. M. and S. McLaughlin. 1993 "Binding of acylated peptides and fatty acids to phospholipid vesicles: pertinence to myristoylated proteins." *Biochemistry* 32(39): 10436-10443.

39. Raju, R. V. et al. 2000 "N-myristoyltransferase." *Mol.Cell Biochem.* 204(1-2): 135-155.

40. Randazzo, P. A. and R. A. Kahn. 1995 "Myristoylation and ADP-ribosylation factor function." *Methods Enzymol.* 250: 394-405.

41. Raulin, J. 2000 "Lipids and retroviruses." *Lipids* 35(2): 123-130.

42. Resh, M. D. 1994 "Myristylation and palmitylation of Src family members: the fats of the matter." *Cell* 76(3): 411-413.

43. Rocque, W. J. et al. 1993 "A comparative analysis of the kinetic mechanism and peptide substrate specificity of human and Saccharomyces cerevisiae myristoyl-CoA:protein N-myristoyltransferase." *J.Biol.Chem.* 268(14): 9964-9971.

44. Schultz, A. M. and S. Oroszlan. 1983 "In vivo modification of retroviral gag gene-encoded polyproteins by myristic acid." *J.Virol.* 46(2): 355-361.

45. Schultz, A. M., E. H. Rabin, and S. Oroszlan. 1979 "Post-translational modification of Rauscher leukemia virus precursor polyproteins encoded by the gag gene." *J.Virol.* 30(1): 255-266.

46. Sikorski, J. A. et al. 1997 "Selective peptidic and peptidomimetic inhibitors of Candida albicans myristoylCoA: protein N-myristoyltransferase: a new approach to antifungal therapy." *Biopolymers* 43(1): 43-71.

47. Slightom, J. L., A. E. Blechl, and O. Smithies. 1980 "Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes." *Cell* 21(3): 627-638.

48. Sunyaev, S. R. et al. 1999 "PSIC: profile extraction from sequence alignments with position- specific counts of independent observations." *Protein Eng* 12(5): 387-394.

49. Towler, D. A. et al. 1988a "Myristoyl CoA:protein N-myristoyltransferase activities from rat liver and yeast possess overlapping yet distinct peptide substrate specificities." *J.Biol.Chem.* 263(4): 1784-1790.

50. Towler, D. A. et al. 1987 "Amino-terminal processing of proteins by N-myristoylation. Substrate specificity of N-myristoyl transferase." *J.Biol.Chem.* 262(3): 1030-1036.

51. Towler, D. A. et al. 1988b "The biology and enzymology of eukaryotic protein acylation." *Annu.Rev.Biochem.* 57: 69-99.

52. Tsunasawa, S., J. W. Stewart, and F. Sherman. 1985 "Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase." *J.Biol.Chem.* 260(9): 5382-5391.

53. Vandekerckhove, J., A. A. Lal, and E. D. Korn. 1984 "Amino acid sequence of Acanthamoeba actin." *J.Mol.Biol.* 172(1): 141-147.

54. Weston, S. A. et al. 1998 "Crystal structure of the anti-fungal target N-myristoyl transferase." *Nat.Struct.Biol.* 5(3): 213-221.

55. Wilcox, C., J. S. Hu, and E. N. Olson. 1987 "Acylation of proteins with myristic acid occurs cotranslationally." *Science* 238(4831): 1275-1278.

56. Zha, J. et al. 2000 "Posttranslational N-myristoylation of BID as a molecular switch for targeting mitochondria and apoptosis [In Process Citation]." *Science* 290(5497): 1761-1765.

57. Zhou, W. and M. D. Resh. 1996 "Differential membrane binding of the human immunodeficiency virus type 1 matrix protein." *J.Virol.* 70(12): 8540-8548.

58. Zozulya, S., D. Ladant, and L. Stryer. 1995 "Expression and characterization of calcium-myristoyl switch proteins." *Methods Enzymol.* 250: 383-393.

Claims

1. A method for identifying proteins with N-terminal N-myristoylation from the knowledge of their amino acid sequence, comprising the steps of

   (a) computational analysis of the N-terminal segment of the primary structure of the query protein and deciding the candidate's status by applying a scoring system based on

      (i) sensitive profile extraction,

      (ii) physical property requirements,

      (iii) compensatory effects over multiple positions, and

   (b) determining by experimental verification whether a protein selected in a) is a substrate for myristoyl-CoA:protein N-myristoyltransferase.

2. The method of claim 1, wherein the scores obtained in step a) are translated into probabilities of false-positive prediction.