J Math Chem
DOI 10.1007/s10910-012-9983-z

ORIGINAL PAPER

Mass distributions of linear chain polymers

Shane L. Hubler · Gheorghe Craciun

Received: 3 November 2011 / Accepted: 14 February 2012

Abstract Biochemistry has many examples of linear chain polymers, i.e., molecules formed from a sequence of units from a finite set of possibilities; examples include proteins, RNA, single-stranded DNA, and paired DNA. In the field of mass spectrometry, it is useful to consider the idea of weighted alphabets, with a word inheriting weight from its letters. We describe the distribution of the mass of these words in terms of a simple recurrence relation, the general solution to that relation, and a canonical form that explicitly describes both the exponential form of this distribution and its periodic features, thus explaining a wave pattern that has been observed in protein mass databases. Further, we show that a pure exponential term dominates the distribution and that there is exactly one such purely exponential term. Finally, we illustrate the use of this theorem by describing a formula for the integer mass distribution of peptides and we compare our theoretical results with mass distributions of human and yeast peptides.

Keywords Linear chain polymers · Peptides · Mass distribution · Mass spectrometry · Protein database

S. L. Hubler (✉)
Biotechnology Center, University of Wisconsin, Madison, WI, USA
e-mail: slhubler@wisc.edu

G. Craciun (✉)
Department of Mathematics, University of Wisconsin, Madison, WI, USA
e-mail: craciun@math.wisc.edu

G. Craciun
Department of Biomolecular Chemistry, University of Wisconsin, Madison, WI, USA

Published online: 01 March 2012

1 Introduction

In chemistry and biology there are many examples of linear chain polymers, i.e., molecules formed from a sequence of units from a finite set of possibilities. Examples include peptides/proteins, RNA, single-stranded DNA, and paired DNA. We want to characterize patterns of the masses of such objects, considering an alphabet of building blocks (e.g. amino acids, in the case of peptides and proteins) and an assigned mass function. Because the molecule is linear, the sequence of building blocks can be thought of as a word over the alphabet.

We are particularly interested in the number $C(M)$ of words that have a particular weight M. Ideally we would like to solve this problem for arbitrary masses but, as we shall see, we will restrict most of our analysis to integer masses.

However, we describe a recurrence relation that is applicable to real-valued masses:

$$C(M) = \sum_{j=1}^{d} C(M - m_j) \tag{1}$$

where d is the number of objects in our alphabet and m_j is the mass of the jth object. In addition, by rescaling the masses, we can generalize all of our results to the case of rational masses.

Next we solve this recurrence relation. The general solution is given by the following theorem:

Theorem 1 (General Solution to Sequence Counting Problem) *Consider a finite set of masses $m_1, \ldots, m_d$ and denote by $C(M)$ the number of sequences of masses (i.e. "words") that have a total mass of M. Then C satisfies a mass recurrence relation (Eq. 1). Furthermore, if the characteristic polynomial of this recurrence relation has distinct roots then there exist real constants $k, c_0, \ldots, c_k, r_0, \ldots, r_k, \theta_1, \ldots, \theta_k, \varphi_1, \ldots, \varphi_k$ such that*

$$C(M) = c_0 r_0^M + \sum_{j=1}^{k} c_j r_j^M \cos(2\pi\theta_j M + \varphi_j) \tag{2}$$

and $r_0 \geq r_j > 0$ for $j = 1, \ldots, k$.

The characteristic polynomial of a recurrence relation will be discussed in the next section. The rest of the paper is devoted to the proof of this theorem and related results, including a generalization for the case where the characteristic polynomial does *not* have distinct roots.

Our interest in this problem originated from the study of mass distributions of peptide fragment ions [1] obtained using mass spectrometry [2,3]. In particular, we (and others [4]) investigated the mass distribution of peptides for both biological and theoretical peptides [5]. This led to the observation that peptide density reaches a local maximum approximately every 14 Da [4,5]. The results described in this paper explain the source of this pattern, predict that periodic patterns are present for all non-trivial

linear chain polymers, and describe the method by which all periodic terms can be enumerated, given a fixed size mass unit (e.g. Daltons).

Finally, we note that computing the general form of the solution involving irrational masses is a very interesting open problem.

2 Definitions and notations

We define an *alphabet* $\mathcal{A}$ as a finite ordered list of objects $\{a_1, \ldots, a_d\}$.

Suppose $\mathcal{A}$ is an alphabet. Then mass$(\cdot)$ is a *mass function* of $\mathcal{A}$ if it is a function from $\mathcal{A}$ to the positive real numbers.

Suppose $\mathcal{A}$ is an alphabet. Then mass$(\cdot)$ is an *integer mass function* of $\mathcal{A}$ if it is a function from $\mathcal{A}$ to the positive integers.

Suppose $\mathcal{A}$ is an alphabet and mass$(\cdot)$ is a mass function of $\mathcal{A}$. Then the pair $\mathcal{O} = (\mathcal{A}, \text{mass}(\cdot))$ is a *collection of mass objects*.

Suppose $\mathcal{O} = (\mathcal{A}, \text{mass}(\cdot))$ is a collection of mass objects with alphabet $\mathcal{A} = \{a_1, \ldots, a_d\}$. Then $m_1, \ldots, m_d$ are the *object masses* of $\mathcal{O}$ if $m_j = \text{mass}(a_j)$ $(j = 1, \ldots, d)$.

Suppose $m_1, \ldots, m_d$ are the object masses of the collection of mass objects $\mathcal{O}$. Then Eq. 1 is the *mass recurrence relation* for $\mathcal{O}$.

Suppose also that $m_1 \leq m_2 \leq \cdots \leq m_d$ are the object masses of the collection of mass objects $\mathcal{O}$. Then p is the *characteristic polynomial* of the recurrence relation for $\mathcal{O}$ if $p(z) = z^{m_d} - \sum_{j=1}^{d} z^{m_d - m_j}$.

Note that this definition of characteristic polynomial is the standard one for linear difference equations of one variable.

Suppose $\mathcal{O} = (\mathcal{A}, \text{mass}(\cdot))$ is a collection of mass objects. Then s is an *ordered sequence* (or simply *sequence*) over $\mathcal{O}$ (or, equivalently, over $\mathcal{A}$) if s is an ordered set of letters, $a_1, a_2, \ldots, a_\ell$ where $a_i \in \mathcal{A}$, $i = 1, 2, \ldots, \ell$. Intuitively, we can think of a sequence over $\mathcal{O}$ as a word formed using letters from the alphabet $\mathcal{A}$.

Suppose $\mathcal{O} = (\mathcal{A}, \text{mass}(\cdot))$ is a collection of mass objects and $s = a_1, a_2 \ldots a_\ell$ is a sequence. The *mass of the sequence* s is $\text{mass}(s) = \sum_{i=1}^{\text{length}}(s)\text{mass}(a_i)$

Suppose $\mathcal{O}$ is a collection of mass objects. Then C is the *sequence counting function* for $\mathcal{O}$ if $C(M) = \#\{\text{sequence } s \text{ over } \mathcal{O} | \text{mass}(s) = M\}$, the number of sequences s over $\mathcal{O}$ where the mass of s is M.

3 Mass recurrence relation

In this section we describe a recurrence relation for the sequence counting function.

Proposition 2 (Mass recurrence relation) *Suppose C is the sequence counting function for a collection of mass objects $\mathcal{O}$ whose masses are $m_1 \leq m_2 \leq \cdots \leq m_k$. Then C satisfies the following mass recurrence relation of $\mathcal{O}$ with initial conditions:*

$$C(M) = \begin{cases} 0 & \text{when } -m_k < M < 0 \\ 1 & \text{when } M = 0 \\ \sum_{j=1}^{d} C(M - m_j) & \text{when } M > 0 \end{cases} \tag{3}$$

Proof Assuming $M > 0$, we start with the definition for a sequence counting function and consider the last letter of every sequence:

$$\begin{aligned} C(M) &= \#\{\text{sequence } s \text{ over } \mathcal{O} | \text{mass}(\mathcal{O}, s) = M\} \\ &= \sum_{j=1}^{d} \#\{\text{sequence } s \text{ over } \mathcal{O} | \text{mass}(\mathcal{O}, s) = M \text{ and the last letter of } s \text{ is } a_j \in \mathcal{A}\} \\ &= \sum_{j=1}^{d} \#\{\text{sequence } s \text{ over } \mathcal{O} | \text{mass}(\mathcal{O}, s) = M - m_j\} \\ &= \sum_{j=1}^{d} C(M - m_j) \end{aligned}$$

The last equality only holds for all $M > 0$ if we define C to be zero for the relevant negative masses that arise. The initial condition related to mass 0 follows from the idea that there is exactly one string of mass zero, namely the null string, ϕ. □

4 Solving the recurrence relation

We next want to solve the recurrence relation. Note that while the recurrence relation described in the previous theorem is applicable for real-valued (or even vector-valued) masses, from now on we require *integer* masses.

Theorem 3 (General solution to a linear homogenous difference equation) *Suppose C satisfies the linear homogenous difference equation*

$$a_n C(M+n) + a_{n-1} C(M+n-1) + \cdots + a_0 C(M+0) = 0 \tag{4}$$

for all positive integers M. Let $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 x^0$ be the characteristic polynomial for the difference equation. Suppose further that p has n distinct roots, $z_1, \ldots, z_n$. Then for any solution of Eq. 4 there exists a unique list of complex numbers $\lambda_1, \ldots, \lambda_n$ such that for all positive integers M,

$$C(M) = \sum_{j=1}^{n} \lambda_j z_j^M \tag{5}$$

Proof This is a key result in many texts on difference equations and a proof for it can be found in, for example, [6–8]. Here is a sketch of the proof: for any root z of p, the sequence z^M satisfies Eq. 4. Therefore all linear combinations $\sum_{j=1}^{n} \lambda_j z_j^M$ satisfy Eq. 4. This form represents the general solution because $(z_j^M)_{\substack{a+1 \le M \le a+n \\ 1 \le j \le n}}$ is a Vandermonde matrix and is, therefore, invertible, giving us a unique solution. □

The next lemma applies Theorem 3 (General solution to a linear homogenous difference equation) to sequence counting functions.

Lemma 4 (Solving the recurrence relation—distinct roots) *Let C be the sequence counting function for a collection of mass objects with integer masses,$\mathcal{O}$, and let p be the characteristic polynomial of the mass recurrence relation of $\mathcal{O}$. Suppose p is of degree n and has n distinct roots, $z_1, \ldots, z_n$. Then there exists a unique set of complex numbers $\lambda_1, \ldots, \lambda_n$ such that*

$$C(M) = \sum_{j=1}^{n} \lambda_j z_j^M.$$

Proof By Proposition 2 (Mass recurrence relation), C satisfies the mass recurrence relation with initial conditions

$$C(M) = \begin{cases} 0 \text{ when } M \in (-n, 0) \\ 1 \text{ when } M = 0 \end{cases}$$

(where n is the largest mass). Note that these initial conditions together with the mass recurrence relation uniquely determine the function C on all positive integers. The mass recurrence relation is a linear homogeneous difference equation with characteristic polynomial p. Because p has distinct roots, we may apply Theorem 3 (General solution to a linear homogenous difference equation) to obtain the desired conclusion. □

Another form of this lemma which handles non-distinct roots can be found in [8].

5 Analysis of the solution of the recurrence relation

Now that we have a solution to the recurrence relation, we want to describe the form of the solution in a way that is easy to interpret. In particular, we describe an exponential growth function overlaid with a periodic pattern. A pattern of this sort has been described in [5,9].

Lemma 5 (Eigenvalues of conjugate pairs are conjugates) *Let C be the sequence counting function for a collection of mass objects $\mathcal{O}$ and let p be the characteristic polynomial of the mass recurrence relation of $\mathcal{O}$. Assume that p has distinct roots, $z_1, \ldots, z_d$ and that $C(M) = \sum_{j=1}^{d} \lambda_j z_j^M$. Suppose that $z_j = \overline{z_{j'}}$. Then $\lambda_j = \overline{\lambda_{j'}}$.*

Proof Note that the conjugate of every root of p is also a root of p. Let j' denote the index such that $z_j = \overline{z_{j'}}$ for each j. (This is well-defined because p has distinct roots.) Then, combined with the fact that $C(M)$ is a real number, we have

$$\sum_{j=1}^{d} \lambda_j z_j^M = C(M) = \overline{C(M)} = \sum_{j=1}^{d} \overline{\lambda_j} \overline{z_j^M} = \sum_{j'=1}^{d} \overline{\lambda_j} z_{j'}^M = \sum_{j=1}^{d} \overline{\lambda_{j'}} z_j^M$$

However, by Lemma 4 (Solving the recurrence relation—distinct roots), we know that the set of $\lambda_1, \ldots, \lambda_d$ is unique. Therefore, $\lambda_j = \overline{\lambda_{j'}}$. □

To simplify the statement of the next theorem, we need the following definition: *We say that that* c_j, θ_j, and φ_j *satisfy standard conditions for* $j \in J$ *if, for all* $j \in J$ we have:

$c_j > 0$,
$0 < \theta_j \leq \frac{1}{2}$, and
$-\pi < \varphi_j \leq \pi$.

The terms c_j, θ_j, and φ_j refer to the magnitude, period, and phase of periodic terms, respectively, defined in the theorem below and used in later statements.

Theorem 6 (Explicit description of sequence counting function—simple form) *Let* C *be the sequence counting function for a collection of mass objects* $\mathcal{O}$ *and* p *be the characteristic polynomial of the mass recurrence relation of* $\mathcal{O}$*. Suppose* p *has distinct roots. Then there exists real numbers* c_j, θ_j, φ_j*, that satisfy the standard conditions for* $1 \leq j \leq k$ *and there exist positive* $c_0, r_0, r_1, \ldots, r_k$ *such that*

$$C(M) = c_0 r_0^M + \sum_{i=1}^{k} c_i r_i^M \cos(2\pi\theta_i M + \varphi_i),$$

and

$$0 < r_j \text{ for all } j = 0, \ldots, k.$$

Proof By Proposition 2 (Mass recurrence relation), we know that C satisfies the recurrence relation. Thus, we have changed the problem to one of solving the recurrence relation, which is a linear difference equation. Let n be the order of p and $z_1, \ldots, z_n$ be the roots of p. By Lemma 4 (Solving the recurrence relation—distinct roots) there exist complex numbers $\lambda_1, \ldots, \lambda_n$, such that $C(M) = \sum_{j=1}^{n} \lambda_j z_j^M$. Note that 0 is not a root since $p(0) \geq 1$. Therefore, we may separate the list of roots into four sets: positive roots R^+, negative roots R^-, complex roots in the upper half-plane R^u, and complex roots in the lower half-plane R^l. We can now rewrite the expression of $C(M)$ as

$$C(M) = \sum_{j:z_j \in R^+} \lambda_j z_j^M + \sum_{j:z_j \in R^-} \lambda_j z_j^M + \sum_{j:z_j \in R^u} \lambda_j z_j^M + \sum_{j:z_j \in R^l} \lambda_j z_j^M.$$

Without loss of generality, we assume that the indices of the roots proceed in order from R^+, R^-, R^u, and R^l. We will handle each summand separately.

Positive root:

Note that R^+ has exactly one element by simple application of Descartes' rule of signs. Denote the single term in the corresponding sum as $\lambda_0 z_0^M$; set $r_0 = z_0$ and $c_0 = \lambda_0$. Thus,

$$\sum_{j:z_j \in R^+} \lambda_j z_j^M = c_0 r_0^M$$

Negative roots:

Define b_- as the largest index of an element in R^-. Let $z_j \in R^-$. We set the constants for the jth term as $r_j = |z_j|$, $c_j = |\lambda_j|$, $\theta_j = \frac{1}{2}$, and $\varphi_j = \begin{cases} 0 \text{ if } \lambda_j \geq 0 \\ \pi \text{ if } \lambda_j < 0 \end{cases}$

If we use this definition for an individual term in the summand we can rewrite the summand as find

$$\begin{aligned} \lambda_j z_j^M &= \lambda_j (-r_j)^M \\ &= c_j r_j^M \cos(2\pi\theta_j M + \varphi_j) \end{aligned}$$

Thus, the summand is

$$\begin{aligned} \sum_{j|z_j \in R^-} \lambda_j z_j^M &= \sum_{i=1}^{b_-} \lambda_j (-r_j)^M \\ &= \sum_{i=1}^{b_-} c_i r_j^M \cos(2\pi\theta_i M + \varphi_i) \end{aligned}$$

Complex roots (upper and lower half-plane):

Define b_u as the largest index j of an element in R^u. Let $z_j \in R^u$. Note that, because z_j is a root of p and p is a polynomial over real numbers, the conjugate of z_j, $\overline{z_j}$, is also a root of p. Call the index of the conjugate root j'; i.e. $z_j = \overline{z_{j'}}$. Note also that $\lambda_j = \overline{\lambda_{j'}}$, by Lemma 5 (Eigenvalues of conjugate pairs are conjugates). Set $c_j = 2|\lambda_j|$ and set φ_j such that $\lambda_j = \frac{1}{2} c_j e^{i\varphi_j}$ and $-\pi < \varphi_j < \pi$. Also set $r_j = |z_j| = |z_{j'}|$ and set θ_j so that $z_j = r_j e^{i2\pi\theta_j}$ and $0 < 2\pi\theta_j < \pi$. In other words, $0 < \theta_j < \frac{1}{2}$. Now we combine the term j from the first complex sum with the term j' from the second sum to get

$$\begin{aligned} \lambda_j z_j^M + \lambda_{j'} z_{j'}^M &= \lambda_j z_j^M + \overline{\lambda_j z_j^M} \\ &= \frac{1}{2} c_j e^{i\varphi_j} r_j^M e^{i2\pi\theta_j M} + \frac{1}{2} c_j e^{-i\varphi_j} r_j^M e^{-i2\pi\theta_j M} \\ &= \frac{1}{2} c_j r_j^M \left(e^{i(2\pi\theta_j M + \varphi_j)} + e^{-i(2\pi\theta_j M + \varphi_j)} \right) \\ &= c_j r_j^M \cos(2\pi\theta_j M + \varphi_j) \end{aligned}$$

Thus, the contribution from the complex roots (those that are not real) is

$$\sum_{j|z_j \in R^u} \lambda_j z_j^M + \sum_{j|z_j \in R^l} \lambda_j z_j^M = \sum_{j=b_-+1}^{b_u} c_i r_j^M \cos(2\pi\theta_i M + \varphi_i).$$

All roots:

Combining the three parts we get

$$\begin{aligned} C(M) &= \sum_{j|z_j \in R^+} \lambda_j z_j^M + \sum_{j|z_j \in R^-} \lambda_j z_j^M + \sum_{j|z_j \in R^u} \lambda_j z_j^M + \sum_{j|z_j \in R^l} \lambda_j z_j^M \\ &= c_0 r_0^M + \sum_{j=1}^{b_-} c_i r_j^M \cos(2\pi\theta_i M + \varphi_i) + \sum_{j=b_-+1}^{b_u} c_i r_j^M \cos(2\pi\theta_i M + \varphi_i) \\ &= c_0 r_0^M + \sum_{j=1}^{k} c_i r_j^M \cos(2\pi\theta_i M + \varphi_i), \end{aligned}$$

where $k = b_u$.

To complete this proof, we need to verify that all of the constants are in the correct ranges; i.e. that they satisfy the standard conditions. The c_j's ($j > 0$) are positive by definition, as are the r_j's. In addition, $0 < \theta_j \leq \frac{1}{2}$ and $-\pi < \varphi_j \leq \pi$ by definition, for $j = 1, 2, \ldots, b_u$. Also, combining the above statements with Lemma 13 (Constant c0 > 0) we know that $c_0 > 0$. □

6 Proof of maximality

The only difference between Theorem 6 (Explicit description of sequence counting function—simple form), above, and Theorem 1 (General Solution to Sequence Counting Problem) is that the latter also tells us that $r_0 \geq r_j$ for all $j \geq 1$. This is the topic of Theorem 14 (Maximality of positive root), found at the end of this section. The rest of this section contains technical lemmas for use in the proof. The intuition of the proof is that if the periodic terms dominate the real term then there must be a point at which the result is negative, which is a contradiction of the fact that the function is a counting function (i.e., must map non-negative integers to non-negative integers).

First, we show that the sum of cosines with rational period over a special set of indices is zero.

Lemma 7 (Sum of rational period cosine is zero) *Let θ be a positive rational number given in reduced form by (a/b). Further, assume that $b > 2$. Let k and w be natural numbers and φ a real number. Then*

$$\sum_{m=w+1}^{w+kb} \cos(2\pi\theta m + \varphi) = 0.$$

Proof First, fix integers k and w and real number φ. Now note that a and b are relatively prime. This means that for every value $j = 0, 1, \ldots, b-1$, there exists an $m \in [w+1, w+b]$ such that $am \equiv j \bmod b$. Therefore,

$$\begin{aligned}\sum_{m=w+1}^{w+b} \cos(2\pi\theta m+\varphi) &= \sum_{m=w+1}^{w+b} \mathrm{Re}(e^{2\pi i\theta m+\varphi}) \\ &= \mathrm{Re}\left(\sum_{m=w+1}^{w+b} e^{2\pi i\frac{a}{b}m+\varphi}\right) \\ &= \mathrm{Re}\left(e^{w+\varphi}\sum_{m=1}^{b} e^{2\pi i\frac{m}{b}}\right)\end{aligned}$$

Note that the last term is the sum of the bth roots of unity. Let $\omega_1, \ldots, \omega_b$ be the bth roots of unity. We want to show that the sum of the bth roots of unity is always zero. Note that these roots are all of the solutions to $x^b - 1 = 0$. Thus,

$$\prod_{j=1}^{b}(x-\omega_j) = x^b - 1$$

Of particular interest, however, is the term of order $b-1$: its coefficient is the sum of all of the roots. Since this coefficient is zero, then the sum of the roots is zero.

Finally, to prove our lemma, we need only add several terms that add to zero:

$$\begin{aligned}\sum_{m=w+1}^{w+kb} \cos(2\pi\theta m+\varphi) &= \sum_{m=w+1}^{w+b} \cos(2\pi\theta m+\varphi) + \sum_{m=(w+b)+1}^{w+2b} \cos(2\pi\theta m+\varphi) \\ &\quad + \cdots + \sum_{m=(w+(k-1)b)+1}^{w+kb} \cos(2\pi\theta m+\varphi) \\ &= 0\end{aligned}$$

□

For simplification in the rest of this section we need to add additional notation. We denote the set of indexes of roots of the characteristic polynomial that have magnitude r, $\{j|r_j = r\}$, *as* $\mathcal{M}_r$. Also, we denote the periodic contribution of the ith root towards $C(M)$, $c_i\cos(2\pi\theta_i M+\varphi_i)$, as $t_i(M)$. We refer to $t_i(M)$ as a parameterized cosine function.

The following lemma states that the sum of $t_i(M)$ repeats a particular negative value infinitely often, provided the $\mathcal{M}_r$ contains only rational values.

Lemma 8 (Functions with rational period repeat negative values) *Suppose θ_j is rational for every $j \in \mathcal{M}_r$ and $f(M) = \sum_{j\in\mathcal{M}_r} c_j\cos(2\pi\theta_j M+\varphi_j)$. Then either f is identically zero on the integers or there exists $\delta > 0$ and integers w and ℓ such that $f(w+k\ell) = -\delta$ for $k = 1, 2, \ldots$*

Proof Since θ_j is rational, it can be represented by $\frac{a_j}{b_j}$ (assume that this is reduced form). Let $\ell = lcm\{b_j|j \in \mathcal{M}_r\}$, the least common multiple of the denominators

of the cosine frequencies. First note that, because $\theta_j \ell$ is an integer and $t_j(M) = c_j \cos(2\pi\theta_j M + \varphi_j)$,

$$\begin{aligned} f(M+\ell) &= \sum_{j\in\mathcal{M}_r} t_j(M+\ell) \\ &= \sum_{j\in\mathcal{M}_r} c_j \cos(2\pi\theta_j M + 2\pi\theta_j \ell + \varphi_j) \\ &= \sum_{j\in\mathcal{M}_r} c_j \cos(2\pi\theta_j M + \varphi_j) \\ &= f(M) \end{aligned}$$

In other words, f is periodic with a period that divides ℓ.

Now, by Lemma 7 (Sum of rational period cosine is zero) we know that, for all integers w and k

$$\sum_{M=w+1}^{w+kb_i} t_j(M) = 0.$$

In particular, because b_j divides ℓ, we know that for all j,

$$\sum_{M=w+1}^{w+\ell} t_j(M) = 0$$

Therefore,

$$\begin{aligned} \sum_{M=w+1}^{w+\ell} f(M) &= \sum_{M=w+1}^{w+\ell} \sum_{j\in\mathcal{M}_r} t_j(M) \\ &= \sum_{j\in\mathcal{M}_r} \sum_{M=w+1}^{w+\ell} t_j(M) \\ &= 0 \end{aligned}$$

If f is identically zero on the integers then we are done. Otherwise, it follows that there exists an integer, w, for which $f(w) = -\delta < 0$. However, f is periodic with a period that divides ℓ. Therefore we know that $w+\ell, w+2\ell, \ldots$ is an infinite sequence of integers such that $f(w) = -\delta < 0$. □

The following lemma extends the previous one by assuming that the members of $\mathcal{M}_r$ are all related to each other by a complex constant.

Lemma 9 (Sum of rationally related cosines of irrational period is negative infinitely often) *Define* $f(M) = \sum_{j\in\mathcal{M}_r} t_j(M)$, *where* t_j *is defined as previously. Suppose* f *is*

not identically zero on $\mathbb{R}$ *and there exists an irrational number* s *such that for every* $j \in \mathcal{M}_r$, *the ratio* $\frac{\theta_j}{s}$ *is rational. Then there exists an irrational number* α *such that* $f(\frac{x}{\alpha})$ *has period 1. Further, there exists a real number* $\varepsilon > 0$ *and a nonempty open interval* $I \subset [0, 1]$ *such that* $x \in I \Rightarrow f(\frac{x}{\alpha}) < -\varepsilon$.

Proof For every $j \in \mathcal{M}_r$, we define a_j and b_j as the reduced form fraction $\frac{a_j}{b_j} = \frac{\theta_j}{s}$. Let $\ell = \text{lcm}_{j\in\mathcal{M}_r}\{b_j\}$ and choose $\alpha = \frac{s}{\ell}$. Define $g(x) = f(\frac{x}{\alpha})$; we will show that g is periodic with period 1.

$$\begin{aligned} g(x+1) &= f\left(\frac{x}{\alpha} + \frac{1}{\alpha}\right) \\ &= f\left(\frac{x}{\alpha} + \frac{\ell}{s}\right) \\ &= \sum_{j\in\mathcal{M}_r} c_j \cos\left(2\pi\theta_j\left(\frac{x}{\alpha} + \frac{\ell}{s}\right) + \varphi_j\right) \\ &= \sum_{j\in\mathcal{M}_r} c_j \cos\left(2\pi\theta_j\left(\frac{x}{\alpha}\right) + \varphi_j + 2\pi\frac{a_j\ell}{b_j}\right) \end{aligned}$$

Note that $\frac{\ell}{b_j}$ is an integer by the definition of ℓ. Therefore, the term $2\pi\frac{a_j\ell}{b_j}$ is an integral multiple of 2π and can be ignored when inside the cosine function. Thus,

$$\begin{aligned} g(x+1) &= \sum_{j\in\mathcal{M}_r} c_j \cos\left(2\pi\theta_j\left(\frac{x}{\alpha}\right) + \varphi_j\right) \\ &= g(x) \end{aligned}$$

In other words, $f(\frac{x}{\alpha})$ has period 1.

Next we prove the second conclusion, that there exists a real number $\varepsilon > 0$ and a nonempty open interval $I \subset [0, 1]$ such that $x \in I \Rightarrow f(\frac{x}{\alpha}) < -\varepsilon$. Recall that we proved the first conclusion by showing that each cosine term in the sum defining g had an integral number of cycles for x between 0 and 1. This implies

$$\begin{aligned} \int_0^1 g(x)dx &= \sum_{j\in\mathcal{M}_r} \int_0^1 c_j \cos(2\pi\theta_j M + \varphi_j)dx \\ &= 0 \end{aligned}$$

However, because f is not identically zero, neither is g. This fact, combined with a zero integral, means that g must be negative for some $w \in (0, 1)$. Define $\varepsilon = -\frac{1}{2}g(w)$. Being a finite sum of continuous functions, g inherits continuity from the cosines. Therefore, there exists an open interval I containing w such that $x \in I \Rightarrow g(x) < -\varepsilon$, satisfying the second conclusion of the lemma. □

The following theorem will allow us to create a form of the previous lemmas without restrictions on the members of $\mathcal{M}_r$ (see the subsequent lemma).

Theorem 10 (Infinite number of special integers) *If the numbers* $1, \alpha_1, \alpha_2, \ldots, \alpha_k$ *are linearly independent over the rational numbers then for any* k *open non-empty intervals* $I_j \subset [0, 1], \quad j = 1, \ldots, k$, *the set*

$$\Gamma = \{n \in \mathbb{Z}^+ | n\alpha_j \bmod 1 \in I_j;\ j = 1, 2, \ldots, k\} \text{ is infinite.}$$

Proof A more general version of this theorem can be found as Theorem 3.13 of [10], where it is attributed to Hardy and Littlewood [11]. □

Recall that that c_j, θ_j, and φ_j satisfy standard conditions for $j \in J$ if, for all $j \in J$ we have:

$c_j > 0$,
$0 < \theta_j \leq \frac{1}{2}$, and
$-\pi < \varphi_j \leq \pi$.

All of the previous lemmas in this section are used in this lemma. It states that the sum of the $t_i(M)$ is less than a particular negative number infinitely often.

Lemma 11 (Sum is negative infinitely often) *Assume that* c_j, θ_j, and φ_j *satisfy the standard conditions for* $j \in \mathcal{M}_r$. *Define the periodic contribution of* $\mathcal{M}_r$ *towards* $C(M)$ *as*

$$\begin{aligned} f(M) &= \sum_{i \in \mathcal{M}_r} c_i \cos(2\pi\theta_i M + \varphi_i) \\ &= \sum_{i \in \mathcal{M}_r} t_i(M) \end{aligned}$$

Then either f *is identically zero on integers or there exists* $\delta > 0$ *and an infinite sequence of positive integers* $M_1, M_2, \ldots$ *such that* $f(M_\ell) < -\delta$ *for all* $\ell = 1, 2, \ldots$.

Proof It is easier to prove this in three steps:

1. Rational case: $i \in \mathcal{M}_r \Rightarrow \theta_i \in \mathbb{Q}$
2. Irrational case: $i \in \mathcal{M}_r \Rightarrow \theta_i \notin \mathbb{Q}$
3. Mixed case: There exists $i, j \in \mathcal{M}_r$ such that $\theta_i \in \mathbb{Q}$ and $\theta_j \notin \mathbb{Q}$

Case 1 Rational case: $i \in \mathcal{M}_r \Rightarrow \theta_i \in \mathbb{Q}$.

Define f as $f(M) = \sum_{i \in \mathcal{M}_r} c_i \cos(2\pi\theta_i M + \varphi_i)$. By Lemma 8 (Functions with rational period repeat negative values) f is either zero on the integers or there exists $\delta > 0$ and integers w and ℓ such that $f(w + k\ell) = -\delta$ for $k = 1, 2, \ldots$ Note that $\delta/2$ and the sequence of integers $w + \ell,\ w + 2\ell, \ldots$ satisfy the conclusion of the theorem.

Case 2 Irrational case: $i \in \mathcal{M}_r \Rightarrow \theta_i \notin \mathbb{Q}$.

We can decompose $\mathcal{M}_r$ into k equivalence classes, $\mathcal{M}_r^1,\ \mathcal{M}_r^2,\ \ldots,\ \mathcal{M}_r^k$, of indexes representing frequencies that are linear combinations of each other over the rational numbers. Select a single member from each of the equivalence classes: $i_j \in \mathcal{M}_r^j$ and define

$s_j = \theta_{i_j}$. We can also decompose the function f to only sum the terms related to the jth equivalence class: $f_j(M) = \sum_{i \in \mathcal{M}_r^j} t_i(M)$. Without loss of generality, each f_j is not identically zero. This allows us to apply Lemma 9 (Sum of rationally related cosines of irrational period is negative infinitely often) to ascertain the existence of positive numbers $\alpha_1, \alpha_2, \ldots, \alpha_k$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$, and open sub-intervals of $[0, 1]$ $I_1, I_2, \ldots, I_k$ such that $g_j(x) = f_j(\frac{x}{\alpha_j})$ has period 1 and $x \in I_j \Rightarrow f_j(\frac{x}{\alpha_j}) < -\varepsilon_j$. Note also that $1, \alpha_1, \alpha_2, \ldots, \alpha_k$ are linearly independent over rational numbers. This allows us to apply Theorem 10 (Infinite number of special integers) to state that the collection of integers

$\Gamma = \left\{ n \in \mathbb{Z}^+ | n\alpha_j \bmod 1 \in I_j;\ j = 1, 2, \ldots, k \right\}$ is infinite. This list of integers is the set we need in order to prove our claim.

To show this, we start with an $n \in \Gamma$.

$$\begin{aligned} f(n) &= \sum_{j=1}^{k} f_j(n) \\ &= \sum_{j=1}^{k} g_j(n\alpha_j) \end{aligned}$$

By the definition of Γ, our selection of n means that, for each of $j = 1, 2, \ldots,$ $(n\alpha_j \bmod 1) \in I_j$. However, we defined I_j so that $g_j(x) < -\varepsilon_j$ for every $x \in I_j$. In particular, this will be true for $n\alpha_j$. If we now define

$$\delta = \sum_{j=1}^{k} \varepsilon_j,$$

we have the following inequality:

$$\begin{aligned} f(n) &= \sum_{j=1}^{k} g_j(\alpha_j n) \\ &< \sum_{j=1}^{k} -\varepsilon_j \\ &< -\delta \end{aligned}$$

Recall that this holds for every $n \in \Gamma$, an infinite set, thus proving this case.

Case 3 Mixed case: there exists $i, j \in \mathcal{M}_r$ such that $\theta_i \in \mathbb{Q}$ and $\theta_j \notin \mathbb{Q}$.

We decompose $\mathcal{M}_r$ in the same way as in Case 2. However, exactly one of the decompositions includes indices of rational frequencies; call this member of the decomposition $\mathcal{M}_r^0$. We decompose f as well:

$$f(M) = \sum_{j=0}^{k} f_j(M),$$
$$f_j(M) = \sum_{i \in \mathcal{M}_r^j} t_i(M)$$

From Case 1 we know that there exists $\delta_0 > 0$ and integers w_0 and ℓ_0 such that, for every integer n,

$$f_0(w_0 + n\ell_0) = -\delta_0.$$

Now we remove $\mathcal{M}_r^0$ from the set of $\mathcal{M}_r$ (calling the result $\mathcal{M}_r'$) and rescale f:

$$f'(x) = f(w_0 + x\ell_0) - f_0(w_0 + x\ell_0).$$

Because f' only contains cosines of irrational period, it satisfies the conditions of Case 2; we conclude that there exists a positive number ε' and an infinite set, $\Gamma' = \{n' \in \mathbb{Z}^+ | f'(n') < -\varepsilon'\}$

Select an $n' \in \Gamma'$ and define

$$\delta = \varepsilon_0 + \varepsilon'$$

Let $n = w_0 + n'\ell_0$. Then

$$\begin{aligned} f(n) &= f_0(n) + f(n) - f_0(n) \\ &= f_0(n) + f(w_0 + n'\ell_0) - f_0(w_0 + n'\ell_0) \\ &= f_0(n) + f'(n') \\ &< -\varepsilon_0 + -\varepsilon' \\ &< -\delta \end{aligned}$$

This completes our proof of Case 3 and, thus, the lemma. □

Using the previous lemma we are now able to put exponential bounds on the size of negative contributions of Eq. 2.

Lemma 12 (Sum of non-positive terms is negative infinitely often) *Suppose* $f(M) = \sum_{j=1}^{k} c_j r_j^M \cos(2\pi\theta_j M + \varphi_j)$ *where* $c_j, \theta_j,$ and φ_j *satisfy the standard conditions (found in definitions). Define* $r_{\max} = \max_j\{r_j\}$. *Then there exists* $\delta > 0$ *and integers* $M_1 < M_2 < \cdots$ *such that* $f(M_\ell) < -\delta r_{\max}^{M_\ell}$ *for* $\ell = 1, 2, \ldots$

Proof (This lemma is identical to Lemma 11 (Sum is negative infinitely often), except for the exponential term in both the definition of f and the conclusion.)

Recall the notation $\mathcal{M}_r = \{j | r_j = r\}$. If all indices j are in $\mathcal{M}_{r_{\max}}$ then the theorem follows directly from Lemma 11 (Sum is negative infinitely often), after factoring out $r_{\max}$; we will assume, therefore, that there are additional indices.

Define $r_- = \max_{j \notin \mathcal{M}_{r_{\max}}} \{r_j\}$ and $c_- = \max_{j \notin \mathcal{M}_{r_{\max}}} \{c_j\}$ and note that the contribution of all the maximal terms towards f is

$$\sum_{i \in \mathcal{M}_{r_{\max}}} r_i^M t_i(M) = r_{\max}^M \sum_{i \in \mathcal{M}_{r_{\max}}} t_i(M)$$

(where t_i is the parameterized cosine function, as defined previously). By Lemma 11 (Sum is negative infinitely often) we know that there exists a $\delta > 0$ and an infinite sequence of integers, $M_1 < M_2 < \cdots$ such that

$$\sum_{j \in \mathcal{M}_{r_{\max}}} t_j(M_\ell) < -2\delta \quad \ell = 1, 2, \ldots$$

Thus,

$$\begin{aligned} \sum_{j \in \mathcal{M}_{r_{\max}}} r_j^{M_\ell} t_j(M_\ell) = r_{\max}^{M_\ell} \sum_{j \in \mathcal{M}_{r_{\max}}} t_j(M_\ell) \; \ell = 1, 2, \ldots \\ < -2 r_{\max}^{M_\ell} \delta \end{aligned}$$

Therefore, given a real number $u > 0$, all integers $M^* > u$ from the sequence above satisfy $\sum_{i \in \mathcal{M}_{r_{\max}}} r_i^M t_i(M) < -\delta r_{\max}^{M^*}$. We select a particular u:

$$u = \frac{\log\left(\frac{c_-}{\delta}\right)}{\log\left(\frac{r_{\max}}{r_-}\right)} + 1$$

By solving for $(r_{\max})^u$ in the definition of u above we find that

$$\begin{aligned} (r_{\max})^u &= \frac{c_-}{\delta} r_-^{u-1} r_{\max} \\ &> \frac{c_-}{\delta} r_-^u \end{aligned}$$

or

$$1 - \frac{\delta}{c_-} \left(\frac{r_{\max}}{r_-}\right)^u < 0$$

We apply this inequality to get

$$\begin{aligned} f(M^*) &= \sum_{i \notin \mathcal{M}_r} r_i^{M^*} t_i(M^*) + \sum_{i \in \mathcal{M}_r} r_i^{M^*} t_i(M^*) \\ &< c_- r_-^{M^*} - 2\delta r_{\max}^{M^*} \\ &= c_- r_-^{M^*} \left(1 - \frac{\delta}{c_-} \left(\frac{r_{\max}}{r_-}\right)^{M^*}\right) - \delta r_{\max}^{M^*} \end{aligned}$$

$$
\begin{aligned}
&= c_- r_-^{M^*} \left(1 - \frac{\delta}{c_-} \left(\frac{r_{\max}}{r_-}\right)^{M^*-u} \left(\frac{r_{\max}}{r_-}\right)^{u}\right) - \delta r_{\max}^{M^*} \\
&< c_- r_-^{M^*} \left(1 - \frac{\delta}{c_-} \left(\frac{r_{\max}}{r_-}\right)^{u}\right) - \delta r_{\max}^{M^*} \\
&< -\delta r_{\max}^{M^*}
\end{aligned}
$$

□

The following lemma says that the coefficient of the purely exponential term is positive and follows easily from the previous lemma.

Lemma 13 (Constant $c_0 > 0$) *Let C be the sequence counting function for $\mathcal{O}$. Assume that C is not identically zero. Suppose that $C(M) = c_0 r_0^M + \sum_{j=1}^{k} c_j r_j^M \cos(2\pi\theta_j M + \varphi_j)$ and that c_j, θ_j, and φ_j satisfy the standard conditions. Then $c_0 > 0$.*

Proof By Lemma 12 (Sum of non-positive terms is negative infinitely often), we know that there exists an integer M' and a real number $\delta > 0$ such that $\sum_{i=1}^{k} c_i r_i^{M' \cos(2\pi\theta_i M' + \varphi_i)} < -\delta r_{\max}^{M'}$. This implies that

$$
\begin{aligned}
C(M') &= c_0 r_0^{M'} + \sum_{j=1}^{k} c_j r_j^{M'} \cos(2\pi\theta_j M' + \varphi_j) \\
&< c_0 r_0^{M'}
\end{aligned}
$$

In particular, if $c_0 \leq 0$ then $C(M') < 0$. Since C is a counting function, this is a contradiction arising from the assumption that $c_0 \leq 0$. Thus, $c_0 > 0$. □

The following theorem describes the closed form of a sequence counting function. However, this theorem applies to the "simple form"; i.e. the case where the characteristic polynomial has distinct roots. If it does not have distinct roots then the solution involves the product of polynomials and exponentials in M inside the summation, instead of simply exponentials inside the summation.

Theorem 14 (Maximality of positive root) *Let C be the sequence counting function for $\mathcal{O}$. Assume that C is not identically zero. Suppose that $C(M) = c_0 r_0^M + \sum_{j=1}^{k} c_j r_j^M \cos(2\pi\theta_j M + \varphi_j)$ and that c_j, θ_j, and φ_j satisfy the standard conditions (defined earlier) with $c_0 > 0$. Then $r_0 \geq r_j$ for $j = 1, \ldots, k$.*

Proof Assume that this is not the case, that the positive root is not maximal. We will show that this leads to a contradiction, namely that we can find an M such that the counting function $C(M)$ is negative.

Let $r_{\max} = \max_j \{r_j\}$. The statement that the positive root is not maximal is equivalent to stating that the index of the positive root, 0, is not included in $\mathcal{M}_{r_{\max}}$ (notation defined at the beginning of this subsection).

Let

$$f(M) = \sum_{j=1}^{k} c_j r_j^M \cos(2\pi\theta_j M + \varphi_j).$$

Then, by Lemma 12 (Sum of non-positive terms is negative infinitely often) we may conclude that there exists $\delta > 0$ and $M_1 < M_2 < \cdots$ such that $f(M_\ell) < -\delta r_{\max}^{M_\ell}$ for $\ell = 1, 2, \ldots$.

Therefore, given a real number $u > 0$, there exists an $M^* > u$ from our list such that $f(M^*) < -\delta r_{\max}^{M^*}$. We select a particular u:

$$u = \frac{\log\left(\frac{c_0}{\delta}\right)}{\log\left(\frac{r_{\max}}{r_0}\right)} + 1$$

By solving for $r_{\max}^u$ in the definition of u and using the assumption that $\frac{r_{\max}^u}{r_0} > 1$ we find that

$$\begin{aligned} r_{\max}^u &= \frac{c_0}{\delta} r_0^{u-1} r_{\max} \\ &> \frac{c_0}{\delta} r_0^u \end{aligned}$$

or

$$1 - \frac{\delta}{c_0}\left(\frac{r_{\max}}{r_0}\right)^u < 0$$

We apply this inequality to get

$$\begin{aligned} C(M^*) &= c_0 r_0^{M^*} + f(M^*) \\ &< c_0 r_0^{M^*} - \delta r_{\max}^{M^*} \\ &= c_0 r_0^{M^*}\left(1 - \frac{\delta}{c_0}\left(\frac{r_{\max}}{r_0}\right)^{M^*}\right) \\ &= c_0 r_0^{M^*}\left(1 - \frac{\delta}{c_0}\left(\frac{r_{\max}}{r_0}\right)^{M^*-u}\left(\frac{r_{\max}}{r_0}\right)^{u}\right) \\ &< c_0 r_0^{M^*}\left(1 - \frac{\delta}{c_0}\left(\frac{r_{\max}}{r_0}\right)^{u}\right) \\ &< 0 \end{aligned}$$

Since C is a counting function, it must be nonnegative; the assumption that the positive root was not maximal has led to a contradiction. Therefore, the positive root is maximal. □

Finally this leads to the main theorem:

7 Proof of main theorem

Proof of Theorem 1 (General Solution to Sequence Counting Problem) Lemma 4 (Solving the recurrence relation—distinct roots) tells us specifically that C satisfies a recurrence relation. Theorem 6 (Explicit description of sequence counting function—simple form) gives us most of the relationship but Theorem 14 (Maximality of positive root) shows that the single purely exponential term dominates the rest, completing the proof. □

8 Extension

It is possible to remove the restrictions on the recurrence relation (unique roots of the characteristic polynomial). We extend Theorem 3 (General solution to a linear homogenous difference equation), Theorem 6 (Explicit description of sequence counting function—simple form), Lemma 12 (Sum of non-positive terms is negative infinitely often), and Theorem 14 (Maximality of positive root).

Theorem 15 (Extension of General solution to a linear homogenous difference equation) *Suppose C satisfies the linear homogenous difference equation (Eq. 4). For all positive integers M. Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0 x^0$ be the characteristic polynomial for the difference equation. Then for any solution of Eq. 4 there exists a unique list of polynomials $p_1, \ldots, p_n$ such that for all positive integers M,*

$$C(M) = \sum_{j=1}^{n} p_j(M) z_j^M.$$

Proof This extension (and how to find the polynomials) can be found in [6]. □

Theorem 16 (Extension of Theorem 6) *Let C, a function of mass M, count how many sequences of a finite set of masses have a combined mass of M. Then C satisfies a mass recurrence relation. Then there exist real constants $k, c_0, \ldots, c_k, r_0, \ldots, r_k, \theta_1, \ldots, \theta_k, \varphi_1, \ldots, \varphi_k$, satisfying the standard conditions, and polynomials $p_1, \ldots, p_k$, whose highest orders have a coefficient of one, such that*

$$C(M) = c_0 r_0^M + \sum_{j=1}^{k} p_j(M) r_j^M \cos(2\pi\theta_j M + \varphi_j).$$

Proof The proof is nearly identical to that of Theorem 6 (Explicit description of sequence counting function—simple form) except that we need to use Theorem 15 (Extension of General solution to a linear homogenous difference equation) in place of Theorem 3 (General solution to a linear homogenous difference equation) and use slightly modified versions of Lemma 4 (Solving the recurrence relation—distinct roots)and Lemma 5 (Eigenvalues of conjugate pairs are conjugates), substituting polynomials for constants. □

Lemma 17 (Extension of Lemma 12) *Suppose* $f(M) = \sum_{j=1}^{k} c_j p_j(M) r_j^M \cos(2\pi\theta_j M + \varphi_j)$ *where* c_j, θ_j, *and* φ_j *satisfy the standard conditions and* $p_j(M)$ *are polynomials whose highest order term have coefficients of 1. Define* $r_{\max} = \max_j\{r_j\}$. *Define* $n_j = Degree(p_j)$ *and* $n_{\max} = \max_{j \in \mathcal{M}_{r_{\max}}} \{n_j\}$. *Then there exists* $\delta > 0$ *and integers* $M_1 < M_2 < \cdots$ *such that* $f(M_\ell) < -\delta M_\ell^{n_{\max}} r_{\max}^{M_\ell}$ *for* $\ell = 1, 2, \ldots$

Proof Define $c_{\max} = \max_j\{c_j\}$. Using t_i, the parameterized cosine function, as defined previously, we can rewrite f as

$$\begin{aligned} f(M) &= \sum_{i=1}^{k} p_i(M) r_i^M c_i \cos(2\pi\theta_j M + \varphi_j) \\ &= \sum_{i=1}^{k} p_i(M) r_i^M t_i(M) \\ &= \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} M^{n_{\max}} r_i^M t_i(M) + \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} \left(p_i(M) - M^{n_{\max}}\right) r_i^M t_i(M) \\ &\quad + \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i < n_{\max}\}} p_i(M) r_i^M t_i(M) + \sum_{i \notin \mathcal{M}_{r_{\max}}} p_i(M) r_i^M t_i(M) \\ &= f_1(M) + f_2(M) + f_3(M) + f_4(M) \end{aligned}$$

where

$$\begin{aligned} f_1(M) &= \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} M^{n_{\max}} r_i^M t_i(M) \\ f_2(M) &= \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} \left(p_i(M) - M^{n_{\max}}\right) r_i^M t_i(M) \\ f_3(M) &= \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i < n_{\max}\}} p_i(M) r_i^M t_i(M) \\ f_4(M) &= \sum_{i \notin \mathcal{M}_{r_{\max}}} p_i(M) r_i^M t_i(M) \end{aligned}$$

The first three terms all have the same r_i, namely $r_{\max}$, while the last term contains all other r_i terms.

The first term, f_1, is the dominant term and can be rewritten as

$$\begin{aligned} f_1(M) &= \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} M^{n_{\max}} r_i^M t_i(M) \\ &= \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} M^{n_{\max}} r_{\max}^M t_i(M) \end{aligned}$$

$$= M^{n_{\max}} r_{\max}^{M} \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} t_i(M)$$

By Lemma 11 (Sum is negative infinitely often) we know that there exists a $\delta > 0$ and an infinite sequence of integers, $M_1 < M_2 < \cdots$ such that

$$\sum_{j \in \mathcal{M}_{r_{\max}}} t_j(M_\ell) < -4\delta \quad \ell = 1, 2, \ldots$$

Thus,

$$\sum_{j \in \mathcal{M}_{r_{\max}}} M_\ell^{n_{\max}} r_j^{M_\ell} t_j(M_\ell) = M_\ell^{n_{\max}} r_{\max}^{M_\ell} \sum_{j \in \mathcal{M}_{r_{\max}}} t_j(M_\ell) \; \ell = 1, 2, \ldots$$
$$< -4 M_\ell^{n_{\max}} r_{\max}^{M_\ell} \delta$$

Therefore, given a real number $u > 0$, all integers $M^* > u$ from the sequence above satisfy

$$f_1(M^*) = \sum_{i \in \mathcal{M}_{r_{\max}}} r_i^{M^*} t_i(M^*)$$
$$< -4 M^{*n_{\max}} r_{\max}^{M^*} \delta.$$

We now want to select a particular u that makes the non-dominant terms f_2, f_3, f_4, sufficiently small compared to f_1.

First, note that f_2 differs from f_1 only in the polynomial factors $(p_i(M) - M^{n_{\max}})$, which have degree at most $n_{\max} - 1$. Thus, there exists a u_2 such that $(p_i(M) - M^{n_{\max}}) r_i^M < \frac{\delta}{k c_{\max}} M^{n_{\max}} r_{\max}^M$ for all $M > u_2$ and for all $i \in \mathcal{M}_{r_{\max}}$ where $n_i = n_{\max}$. Therefore, for all $M > u_2$,

$$\begin{aligned} f_2(M) &= \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} \left(p_i(M) - M^{n_{\max}}\right) r_i^M t_i(M) \\ &< \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} \left(p_i(M) - M^{n_{\max}}\right) r_i^M c_{\max} \\ &< \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i = n_{\max}\}} \frac{\delta}{k} M^{n_{\max}} r_{\max}^M \\ &< \delta M^{n_{\max}} r_{\max}^M \end{aligned}$$

Similarly, because of the maximality of $n_{\max}$, there exists a u_3 such that

$$f_3(M) = \sum_{\{i \in \mathcal{M}_{r_{\max}} | n_i < n_{\max}\}} p_i(M) r_i^M t_i(M)$$
$$< \delta M^{n_{\max}} r_{\max}^M$$

Finally, because of the maximality of $r_{\max}$, there exists a u_4 such that for all $M > u_3$,

$$\begin{aligned} f_4(M) &= \sum_{i \notin \mathcal{M}_{r_{\max}}} p_i(M) r_i^M t_i(M) \\ &< \delta M^{n_{\max}} r_{\max}^M \end{aligned}$$

We select a $u = \max(u_2, u_3, u_4)$.

We apply the properties of u and we get the following for every $M^* > u$ where $M^* \in \{M_1, M_2, \ldots\}$:

$$\begin{aligned} f(M^*) &= f_1(M^*) + f_1(M^*) + f_1(M^*) + f_1(M^*) \\ &< -4\delta {M^*}^{n_{\max}} r_{\max}^{M^*} + \delta {M^*}^{n_{\max}} r_{\max}^{M^*} + \delta {M^*}^{n_{\max}} r_{\max}^{M^*} + \delta {M^*}^{n_{\max}} r_{\max}^{M^*} \\ &< -\delta {M^*}^{n_{\max}} r_{\max}^{M^*} \end{aligned}$$

□

Theorem 18 (Complete Extension of General Solution to Sequence Counting Problem) *Under the conditions of Theorem 16 (Extension of Theorem 6), $r_0 \geq r_j$ for $j = 1, \ldots, k$.*

Proof Assume that this is not the case, that the positive root is not maximal. We will show that this leads to a contradiction, namely that we can find an M such that the counting function $C(M)$ is negative.

Let $r_{\max} = \max_j \{r_j\}$. The statement that the positive root is not maximal is equivalent to stating that the index of the positive root, 0, is not included in $\mathcal{M}_{r_{\max}}$ (notation defined at the beginning of this subsection).

Let

$$f(M) = \sum_{j=1}^{k} p_j(M) r_j^M c_j \cos(2\pi\theta_j M + \varphi_j).$$

Then, by Lemma 17 (Extension of Lemma 12) we may conclude that there exists $\delta > 0$ and $M_1 < M_2 < \cdots$ such that $f(M_\ell) < -\delta M_\ell^{n_{\max}} r_{\max}^{M_\ell}$ for $\ell = 1, 2, \ldots$

We now choose a particular member M^*, from our list with the property that

$$M^* > \max\left\{1, \frac{\log\left(\frac{c_0}{\delta}\right)}{\log\left(\frac{r_{\max}}{r_0}\right)} + 1\right\}$$

By solving for $r_{\max}^{M^*}$ in the definition of u and using the assumption that $\frac{r_{\max}}{r_0} > 1$ we find that

$$\begin{aligned} r_{\max}^{M^*} &> \frac{c_0}{\delta} r_0^{M^*-1} r_{\max} \\ &> \frac{c_0}{\delta} r_0^{M^*} \end{aligned}$$

or

$$1 - \frac{\delta}{c_0}\left(\frac{r_{\max}}{r_0}\right)^{M^*} < 0$$

We apply this inequality (and the fact that $M^* > 1$) to get

$$\begin{aligned} C(M^*) &= c_0 r_0^{M^*} + f(M^*) \\ &< c_0 r_0^{M^*} - \delta r_{\max}^{M^*} M_{n_{\max}}^* \\ &< c_0 r_0^{M^*} - \delta r_{\max}^{M^*} \\ &= c_0 r_0^{M^*}\left(1 - \frac{\delta}{c_0}\left(\frac{r_{\max}}{r_0}\right)^{M^*}\right) \\ &< 0 \end{aligned}$$

Since C is a counting function, it must be nonnegative; the assumption that the positive root was not maximal has led to a contradiction. Therefore, the positive root is maximal. □

Example 19 (Extension of General Solution is Necessary) Suppose we have an alphabet $\boldsymbol{a} = \{a_1, a_2, a_3, a_4, a_5\}$ and define the masses $m_1 = m_2 = m_3 = 4$ and $m_4 = m_5 = 6$. Then the mass recurrence relation is $C(M) = 3C(M-4) + 2C(M-6)$ and the characteristic polynomial is $p(x) = x^6 - 3x^2 - 2$. The roots of p are $\pm\sqrt{2}$ and $\pm i$ (twice). The explicit form of C is

$$\begin{aligned} C(M) &= c_0\sqrt{2}^M + c_1\left(-\sqrt{2}\right)^M + c_2\cos\left(\frac{\pi}{2}M + \varphi_2\right) + c_3 M\cos\left(\frac{\pi}{2}M + \varphi_3\right) \\ &= c_0\sqrt{2}^M + c_1\left(\sqrt{2}\right)^M\cos(\pi M + 0) + c_2\cos\left(\frac{\pi}{2}M + \varphi_2\right) \\ &\quad + c_3 M\cos\left(\frac{\pi}{2}M + \varphi_3\right). \end{aligned}$$

Note, in particular, the last term, which is linear in M; it is of the form $c_3 q(M) 1^M \cos(\theta M + \varphi)$ where $q(M) = M$, a polynomial of degree 1 in M.

9 Applications

As an application of the above work, we analyze the mass distributions of peptides/proteins. First we recognize that the objects forming the alphabet are the residues of the amino acids. Proposition 2 (Mass recurrence relation) defines the relationship for this special case as

$$\begin{aligned} C(M) = {} & C(M-57) + C(M-71) + C(M-87) + C(M-97) + C(M-99) \\ & + C(M-101) + C(M-103) + 2C(M-113) + C(M-115) \end{aligned}$$

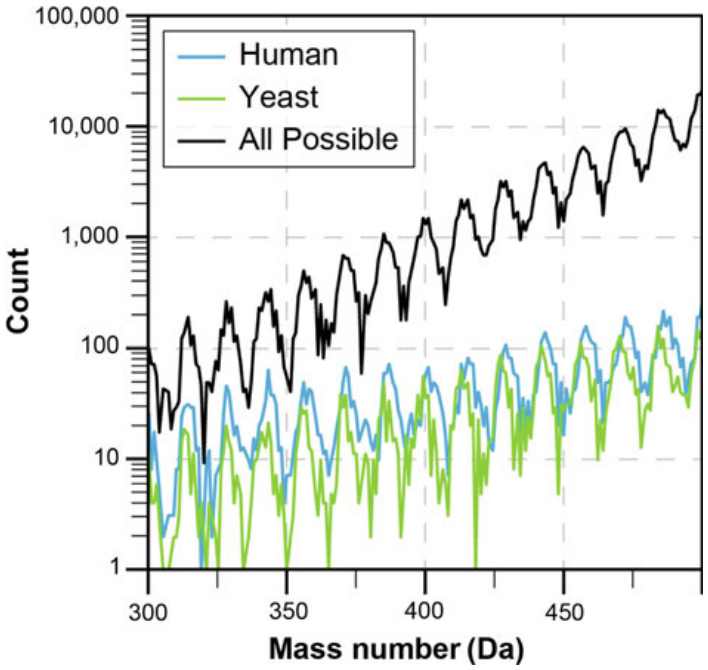

Fig. 1 Comparison of theoretical to empirical mass distributions. The theoretical mass distribution of all peptides in the mass range of 300–500 Da has a wave pattern similar to that of the distribution for human and yeast peptides

$$+C(M-131)+C(M-137)+2C(M-128)+C(M-129)$$
$$+C(M-147)+C(M-156)+C(M-163)+C(M-186),$$

where the numbers 57, 71, 87, . . ., are the masses of amino acid residues measured in Daltons, to the nearest integer. This allows us to analyze the mass distribution of all possible (theoretical) peptides with mass distributions of peptides that appear within a given organism (see Fig. 1).

Next, we determine the characteristic polynomial,

$$x^{186}-x^{129}-x^{115}-x^{99}-x^{89}-x^{87}-x^{85}-x^{83}-2x^{73}$$
$$-x^{72}-x^{71}-2x^{58}-x^{57}-x^{55}-x^{49}-x^{39}-x^{30}-x^{23}-x^{0}$$

We can then approximate the roots (Fig. 2) of the characteristic polynomial using a solver, such as the "roots" function in MATLAB [12]. In this case, as in all we have observed in biological applications, the characteristic polynomial has distinct roots. Therefore, we can apply Theorem 1 (General Solution to Sequence Counting Problem), yielding an exponential term summed with a collection of terms that contain a periodic function multiplied by an exponential function, the former dominating the latter (Fig. 3). If we write the exponential terms in Eq. 2 by powers of 2, then we can easily read the doubling time of each term. Furthermore, if we write the frequencies θ_j in terms of wavelength, then we can emphasize the period of each term. With these conventions, the first three terms are

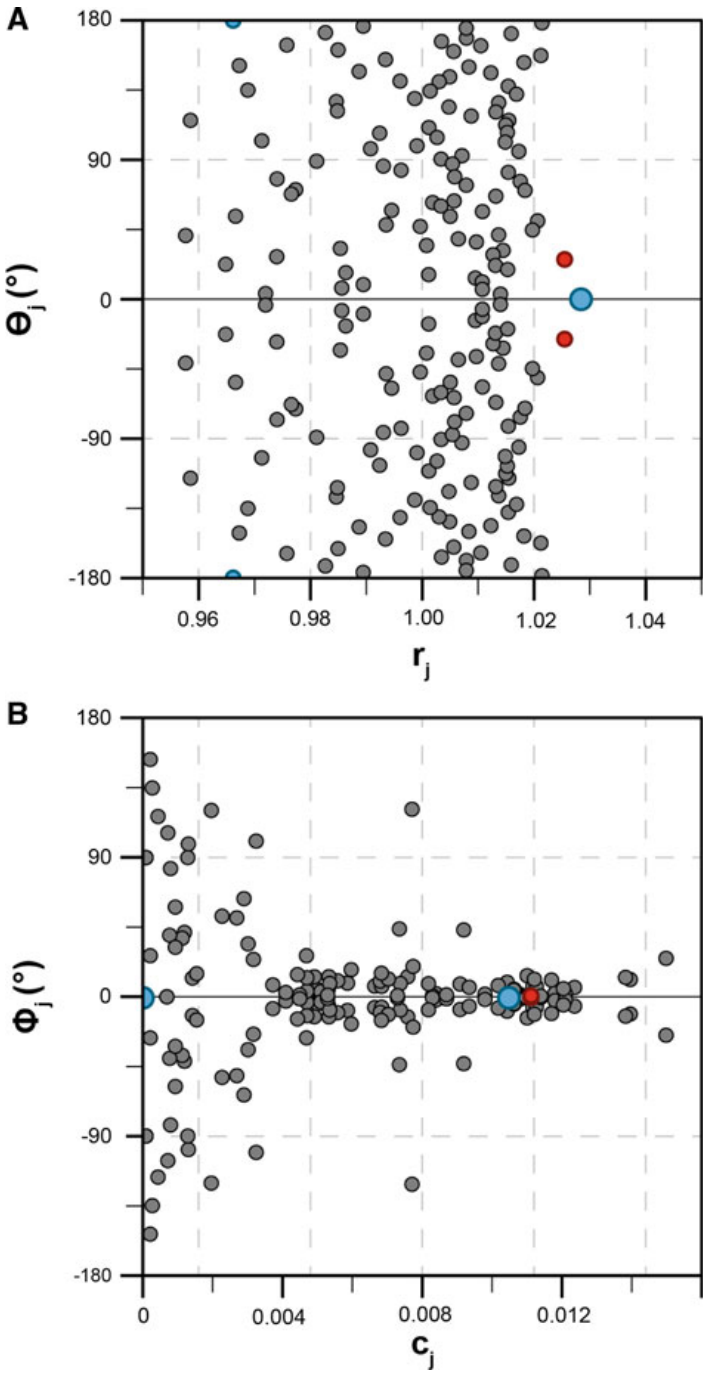

Fig. 2 Roots and eigenvalues of characteristic polynomial for peptides. **a** The roots of the characteristic polynomial for the recurrence relation that uses the integer residue masses of amino acids. As predicted, there is exactly one positive root and it has the largest magnitude (*blue dot*). The *red dots* correspond to the roots of the next largest magnitude; their period is 14.28 Da. **b** The constant terms in Eq. 5 (Color figure online)

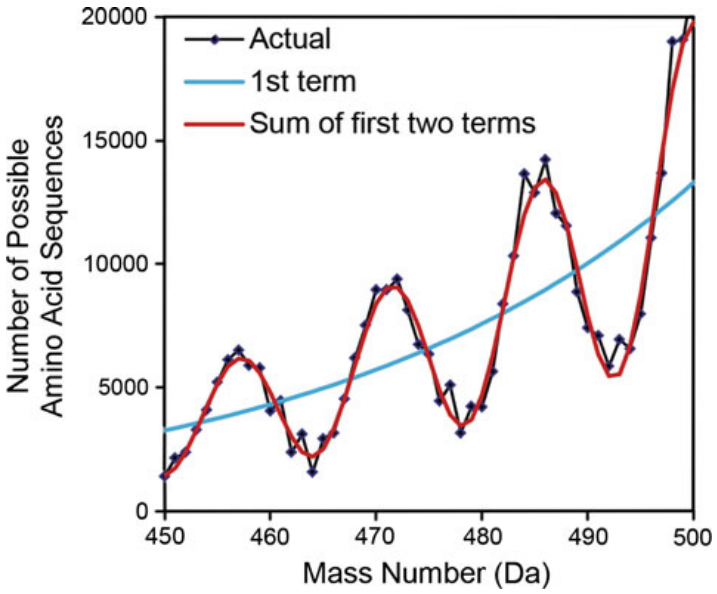

Fig. 3 Comparison of first two terms to actual distribution. The two dominant terms of Eq. 2 for the peptide problem are increasingly accurate as we change the mass region being examined

$$C(M) = \frac{1}{95.238}2^{\frac{M}{24.67}} + \frac{1}{47.04}2^{\frac{M}{27.53}}\cos\left(1.032 + \left(\frac{360}{14.28}\right)M\right) + \frac{1}{45.41}2^{\frac{M}{32.67}}\cos\left(13.676 + \left(\frac{360}{2.023}\right)M\right) + \cdots$$

In total, there is one purely exponential term (which doubles in size every 24.67 Da) and 93 periodic terms we can consider, each of which gives a different periodicity and dominance in the distribution of masses. However, we found that the first few terms are sufficient to approximate the distribution of peptides, at least when rounding the amino acid masses to the nearest Dalton. In particular, the dominant periodic term has a period of 14.28 Da. Solutions for higher mass accuracy are possible by changing units. More examples and details are available in [5].

We expect that these findings could improve estimates of distributions, thus improving a class of algorithms that score peptide identifications in mass spectrometry [13–15], those that require knowledge of the number of possible peptide sequences within a particular mass range.

Acknowledgments SLH was supported by NLM training grant Computation and Informatics in Biology and Medicine 5T15LM007359. GC and SLH were supported by NSF DBI-0701846.

References

1. S.L. Hubler, A. Jue, J. Keith, G.C. McAlister, G. Craciun et al., Valence parity renders z(center dot)-type ions chemically distinct. J. Am. Chem. Soc. **130**, 6388–6394 (2008)

2. M. Scigelova, A. Makarov, Orbitrap mass analyzer—overview and applications in proteomics. Proteomics **6**, 16–21 (2006)
3. JJ Coon, Collisions or electrons? Protein sequence analysis in the 21st century. Anal. Chem. **81**, 3208–3215 (2009)
4. J.D.M.Y. Tipton, C.L. Hendrickson, A.G. Marshall, Utility of the valence parity rule for ECD/AIECD FT-ICR MS: H-dot atom transfer, isobars, and isomers for peptide analysis. Madison, 7–10 December 2008
5. S.L. Hubler, Mathematical analysis of mass spectra data, PhD Thesis, University of Wisconsin-Madison, Madison, 2010
6. V.K. Balakrishnan, *Introductory Discrete Mathematics* (Dover Publications, New York, 1996), vol. xiv, p. 236
7. T.H. Cormen, *Introduction to Algorithms* (The MIT Press, Cambridge, 2009)
8. W.G. Kelley, A.C. Peterson, *Difference Equations: An Introduction with Applications* (Academic Press, Boston, 1991), vol. xi, p. 455
9. S.L. Hubler, G. Craciun, *Periodic Patterns in Distributions of Peptide Masses* (2011)
10. V. Bergelson, in *Minimal idempotents and ergodic Ramsey theory*, ed. by S. Bezuglyi, S. Kolyada Topics in Dynamics and Ergodic Theory (Cambridge University Press, Cambridge, 2003), pp. 8–39
11. G. Hardy, J. Littlewood, Some problems of diophantine approximation (part I). Acta Mathematica **37**, 155–191 (1914)
12. *MATLAB. 7.4 ed.* (The Mathworks Inc., Nattick, 2007)
13. G. Alves, A.Y. Ogurtsov, W.W. Wu, G. Wang, R.F. Shen et al., Calibrating e-values for MS2 database search methods. Biol. Direct 2 (2007)
14. S. Kim, N. Bandeira, P.A. Pevzner, Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. Mol. Cell. Proteom. **8**, 1391–1400 (2009)
15. L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu et al., Open mass spectrometry search algorithm. J. Proteome Res. **3**, 958–964 (2004)