Spring 2015

Mathematical approaches to food nutrient content estimation with a focus on phenylalanine

Jieun Kim
Department of Mathematics, Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Jieun Kim

Entitled
Mathematical Approaches to Food Nutrient Content Estimation with a Focus on Phenylalnine

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Mireille Boutin
Chair

Gregery Buzzard

Nung Kwan Aaron Yip

Carol J. Boushey

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Mireille Boutin

Approved by: David Goldberg 4/23/2015
Head of the Departmental Graduate Program Date

MATHEMATICAL APPROACHES TO FOOD NUTRIENT CONTENT ESTIMATION WITH A FOCUS ON PHENYLALANINE

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Jieun Kim

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

To my beloved family and fiancé John.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor, Professor Mireille Boutin for her continued support, encouragement, and guidance throughout my PhD research. It was my great honor to have had the opportunity to work with and learn from Professor Boutin. Without her insightful advice and persistent help, I would not have been able to complete this dissertation. I would like to sincerely thank my committee members, Professor Carol Boushey, Professor Greg Buzzard, and Professor Aaron Nung Kwan Yip. I also would like to give my special appreciation to Professor Ho-Jong Jang in Hanyang University for inspiring me to pursue this graduate degree. I am deeply thankful for my beautiful family and John, who are always there to encourage me not to give up. Finally, I would like to thank my dear friends for the lovely memories at Purdue for the last six years.

TABLE OF CONTENTS

Page

LIST OF TABLES v

LIST OF FIGURES vi

ABSTRACT vii

1 Introduction 1

2 New multipliers for estimating the phenylalanine content of foods from the protein content 5
2.1 Material and methods 5
2.2 Results and discussion 7
2.2.1 USDA database 7
2.2.2 Danish databank 8
2.3 Conclusions 11

3 Two simple guidelines to limit the phenylalanine intake from sweets with gelatin 13
3.1 Methods 13
3.2 Results and Discussion 15

4 A method for estimating the nutrient content of commercial foods from their label 19
4.1 Step 1: Nutrient content estimation using Approximate ingredient amounts 22
4.1.1 Initial range estimate 22
4.1.2 Iterative method to narrow the range estimate 25
4.2 Step 2: Nutrient content estimate Refinement using Simplex algorithm 28
4.3 Numerical Experiments 30
4.3.1 Convergence of ingredient amounts (A_i) 30
4.3.2 Application to Phenylalanine (Phe) content estimation . . . 31
4.4 Summary and Conclusions 36

5 Conclusion 39

REFERENCES 40

VITA 42

LIST OF TABLES

Table Page

2.1 List of empirical mean and standard deviation for types of food from two sources, USDA database (US), Danish databank (D) and the combined results from two databases (C). A more complete list of Phe:protein ratios can be found in [12]. 10

3.1 Upper bounds on food amounts to limit Phe intake to no more than 20mg 17

4.1 Comparison of phenylalanine content estimates obtained with our methods, two food databases and the full linear programming approach (Simplex Algorithm) (1). 34

4.2 Comparison of phenylalanine content estimates obtained with our methods, two food databases and the full linear programming approach (Simplex Algorithm) (2). 35

LIST OF FIGURES

Figure		Page
2.1	Distribution of Phe:protein ratio for all foods in the USDA database . .	8
2.2	Distribution of Phe:protein ratio for all foods in the Danish databank .	9
4.1	Schematic Diagram of Proposed Method to Estimate the Ingredient Amounts	21
4.2	**Range of Estimates for Ingredient Amounts.** As more nutrients are taken into account, the difference between the estimated maximum amount and the estimated minimum amount for each ingredient often decreases quickly.	32

ABSTRACT

Kim, Jieun PhD, Purdue University, May 2015. Mathematical approaches to food nutrient content estimation with a focus on phenylalanine. Major Professor: Mireille Boutin.

Managing the intake of a certain nutrient can be an effective treatment for some inherited metabolic disorders. An example of such dietary treatments is for phenylketonuria (PKU), for which patients must follow a low-phenylalanine diet for life. Some food databases provide the phenylalanine (Phe) content for a large number of unprocessed foods, and a limited number of composite foods; however, they are not exhaustive. As an attempt to complete this list, we introduce three mathematical approaches to estimate a bound for the Phe content based on the available nutritional information.

The first approach is based on the statistical distribution of the Phe to protein ratios. To be precise, we propose the multipliers 20 and 65 to obtain a minimum bound and a maximum bound for the Phe content from the protein content. The second approach is based on two simple lemmas which apply to sweets with gelatin. Specifically, we show that simple arithmetic operations can be used to determine an amount of sweets that is guaranteed to contain less than 20 mg Phe. The third approach is based on numerical optimization. We use the ingredient list and the Nutrition Facts Label to set up a set of inequalities which we solve numerically. The first step of our solution provides estimates for the ingredient amounts. This can be viewed as an approximate inverse recipe method. Although these mathematical methods are primarily motivated by the problem of estimating the Phe content, they can also be applied to estimating the content of other nutrient. In particular, they could be used to complete missing values in current food composition databases.

1. Introduction

The Nutrition Facts Label on a package of commercial food provides a part of the nutritional information for the food. While this can be a good source for some nutrient content, the Nutrition Facts Label is missing data for many nutrients—for instance, phenylalanine, folic acid, riboflavin, and so on. Furthermore, the precision of listed nutrient content is sometimes too coarse because of large rounding errors. This is problematic when managing strict medical diets. The metabolic disorder phenylketonuria (PKU) is one such example, when individuals should follow a restricted dietary guideline for a nutrient not presented on the food label. Our research is motivated by an attempt to fill the missing data mathematically to support the modified diet.

PKU is characterized by the deficiency of an enzyme called phenylalanine hydroxylase that is necessary to metabolize an amino acid phenylalanine (Phe) into tyrosine. The lack of phenylalanine hydroxylase causes an abnormal accumulation of Phe in an individual's body, which can lead to severe intellectual disabilities [1]. The incidence of PKU is 1 in 15,000 newborn babies in the United States [2]. However, the incidence can differ among regions; for example, 1 in 2,622 newborns from the population in Turkey [3], 1 in 13,290 newborns from the population in British Columbia, Canada [4], and 1 in 12,420 newborns from the population in the West Midlands, UK [5]. The current mainstream treatment for individuals with PKU to maintain low blood Phe level (120-360 μmol/l) is a restricted-Phe diet that should be managed over a lifetime [6]. In order to follow the restricted diet, individuals with PKU should monitor their Phe intake at all times.

One of the biggest concerns that individuals with PKU encounter when managing their restricted-Phe diet is the limited resources of current databases containing information about the presence of phenylalanine. The most widely used database, the USDA Standard Reference Database [7], provides 8194 food items and only a part of

the items (4843) are fully analyzed for amino acid composition. This may be caused by many practical reasons including costly experiments required to measure nutrient content. Therefore, the nutritional information presented by the database is not the complete list, considering the wide range of existing foods. With regard to individuals with PKU, this lack of completeness underestimates the number of foods allowed in their diets.

In order to assist individuals with PKU to broaden their allowable food list, we developed three different mathematical approaches for estimating the unknown Phe content of commercial foods. The three approaches not only evaluate the maximal possible value but also evaluate the minimal possible value for the nutrient content. Because nutritional imbalance can affect development, these individuals must have sufficient protein intake even while maintaining a low-Phe diet. In other words, excluding all proteins will not be an ideal dietary direction, and so the individuals are instructed to supplement their diet by a medical formula. Moreover, limiting protein consumption from an overly exaggerated Phe content will impede the individual's dietary freedom. Thus, estimating the minimal value is in our interest as much as the maximal value. Method 1 yields bounds for the Phe content based on the statistical distribution of Phe to protein ratios. Method 2 suggests an acceptable amount for individuals with PKU to take a specific type of food using a simple calculation. Method 3 utilizes all available information to numerically produce the most compact range of minimum and maximum bounds for the Phe content.

Since Phe is an amino acid, the proportion of Phe in foods greatly depends on the food's proportion of protein. Therefore, deriving the Phe content from protein content has been a longstanding method. The current convention multiplies 50 to the upper bound of the protein content to obtain the rough Phe content of a food [8]. However, according to the statistical distribution in current databases—the USDA Standard Reference Database and the Danish Food Composition Databank (Danish databank [9])—Phe to protein ratios are not always exactly 50; rather, it is between 20 and 65 for more than 97% of the food list. Therefore, it is not safe to use the

conventional multiplier 50 for Phe content estimation. With this logic, Method 1 proposes 20 and 65 as new multipliers in order to accommodate a majority of the cases, which is described in Chapter 2. By using the new multipliers on the protein content, we can suggest the first minimum and maximum bounds for the Phe content of a food.

The second and third approaches use mathematical reasoning and computations to derive a nutrient content from a sorted ingredient list and/or some nutrient data. We assume that no part of any ingredient is removed during the preparation process. Chapter 3 provides a brief explanation of how this approach can be used by considering the problem of estimating the Phe content of sweets made with gelatin and Phe-free ingredients. It is critical for individuals with PKU at early ages to pay extra attention to keeping a low-Phe diet to avoid severe brain damage. Method 2 would help the parents of PKU patients to quickly decide whether they can allow their children a sweet or not, without the fear of exceeding their Phe allowance. When they do not have sufficient time to access a database, they can use the guidelines from Method 2 with only minimal arithmetic (counting a rank of gelatin in the ingredient list and/or dividing a serving size by a maximal bound of the protein content).

This approach is developed and expanded to cases of general commercial foods in Chapter 4. Based on a food label and the USDA Standard Reference Database, Method 3 establishes initial upper and lower bounds for each ingredient amount and refines the bounds iteratively using the properties of inequality. This is an approximate inverse recipe method. Based on these bounds, we approximate a minimum and maximum possible value for the Phe content. This interval of bounds can be further narrowed using a linear programming algorithm such as the Simplex algorithm. To test Method 3, we experimented with 25 commercial foods, the results of which are shown on Table 4.1 and 4.2. In a majority of cases (17/25), the bounds obtained were within 10.4mg of each other, and thus our method provided a very accurate estimate ($\pm$5.2mg) for the Phe content of the foods. We have also created web and Android

applications based on the framework of these mathematical approaches, and these are available to the PKU community at `https://engineering.purdue.edu/brl/PKU/`.

2. New multipliers for estimating the phenylalanine content of foods from the protein content

The protein content of a serving of food is a good indicator of whether the food is appropriate for the PKU diet. Generally, foods containing one or more grams of protein per serving must be carefully measured and the corresponding Phe content consumed must be recorded. This requires knowing the Phe content of the food. A long standing method for getting an estimate for the Phe content of a food consists of multiplying the protein content in grams by 50 in order to get an upper bound on the Phe content (in milligrams) [8, 10]. This method is based on the assumption that Phe constitutes roughly no more than 5% of protein weight. One can somewhat refine this estimate by considering the ingredients. For example, if the food is made of vegetables, then the multiplier 40 is used to estimate the average Phe, while the multiplier 30 is used for fruits. These are long standing conventions, which are still commonly used today for PKU management [8]. In this chapter, we shall test these. More specifically, we shall test the following hypotheses.

1. The Phe:protein ratio in foods is between about 30 mg/g and 50 mg/g.
2. The average Phe:protein ratio in foods made of fruits is about 30 mg/g.
3. The average Phe:protein ratio in foods made of vegetables is about 40 mg/g.

This work was accepted and will be published by the Journal of Food Composition and Analysis [11].

2.1 Material and methods

We studied the statistical distribution of the Phe:protein ratio of foods listed in two food databases, namely the USDA Standard Reference Database (USDA database

[7]), and the Danish Food Composition Databank (Danish databank [9]). The USDA database contains a total of 8194 food items divided into 25 categories, including 'Fruit and Fruit Juices' and 'Vegetables and Vegetable Products'. The protein content is listed for each item rounded to the nearest ten-thousandth of a gram for one gram of a food. It also lists the Phe content rounded to the nearest hundredth of a milligram for a large number of items (4843). In our analysis, we only considered these 4843 food items for which the Phe entry was not empty. We found 99 food items with zero protein content and non-empty Phe content. Four food items were found to have zero protein content but non-zero Phe content. These were excluded. We took all foods with non-zero protein content and divided their Phe content (in milligrams) by their protein content (in grams). The error of the computation for a food having x milligram Phe content with δx standard error and y gram protein content ($y \neq 0$) with δy standard error was calculated as

$$\varepsilon \approx \begin{cases} \frac{\delta x+0.005}{y}, & \textit{if } x = 0, \\ \frac{x}{y}\left(\frac{\delta x+0.005}{x} + \frac{\delta y+0.00005}{y}\right), & \textit{otherwise.} \end{cases}$$

The number 0.005 and 0.00005 are the errors in Phe content and protein content, respectively. These numbers are obtained from the specified rounded decimal places for 100g food items in the USDA database, adjusted for 1g. The Danish databank provides nutritional information for 1050 number of food items divided into 17 categories, including Fruit and fruit products and Vegetables and vegetable products, with protein content rounded to the nearest thousandth. The Phe content of 739 of these items is listed, rounded to the nearest hundredth. We computed the Phe:protein ratio for the Danish databank in the same manner as for the USDA database. However the precision of the protein content in this case (0.0005) yields the following error.

$$\varepsilon \approx \begin{cases} \frac{\delta x+0.005}{y}, & \textit{if } x = 0, \\ \frac{x}{y}\left(\frac{\delta x+0.005}{x} + \frac{\delta y+0.0005}{y}\right), & \textit{otherwise.} \end{cases}$$

2.2 Results and discussion

2.2.1 USDA database

In Figure 2.1, we show a histogram illustrating the distribution of the Phe:protein ratios we obtained from the USDA database. We computed the empirical mean (42.580) and standard deviation (11.469) of all these Phe:protein ratios and plotted the corresponding normal approximation on top of the histogram. Note that the maximum Phe:protein ratio is 546.54 : 1 (for aspartame). The next highest values are for breadfruit seeds (about 108 mg Phe per gram protein) and sweet green peppers (about 107 mg Phe per gram protein), followed by sweet potato chips (about 95 mg Phe per gram protein). Notice the large bump around 40, and another significant one around 50. The errors for the Phe:protein ratios are < 5 in more than 97.132% of cases. Moreover, the errors are < 1 for 71.811% of the foods analyzed in the USDA database. According to our data, the multipliers 30-50 comprise only 76.260% of listed foods with nonnegative Phe content and positive protein content on the USDA database. In other words, the multipliers of Hypothesis 1 are not reliable for 23.740% of the foods. In the USDA database, the maximum and minimum values for Phe:protein ratio are 0 : 1 and 546.54 : 1. However, among 4743 food items with positive protein content, 54 food items have values less than 20 : 1 and 34 food items have values larger than 65 : 1. As shown in Figure 2.1, $20 : 1 \leq$ Phe:protein $\leq 65 : 1$ is true for 98.145% of listed foods. Therefore, we suggest replacing the multipliers 30 and 50 by the new multipliers 20 and 65. Let us now consider the category 'Fruit and Fruit Juices' and the category 'Vegetables and Vegetable Products'. The empirical mean and standard deviation of the Phe:protein ratio for the former category are 30.420 and 12.558. The empirical mean and standard deviation of the Phe:protein ratio for the latter category are 39.522 and 11.432. There are 10 food items in the 'Fruit and Fruit Juices' category which have a Phe:protein ratio around 60. Nine of these are grapefruit products, which according to our communication with USDA-ARS Nutrient Data Laboratory, are erroneous entries which should be adjusted to

the value of 'Grapefruit, raw, pink and red, all areas'(16.883 ± 1.461) . A total of 12 items are listed as containing zero milligram of Phe but non-zero protein content. These include thousand island Salad dressing, mayonnaise, and cream cheese-flavor Frostings, as well as 'Alcoholic beverage, beer', both 'regular' and 'Light'. The latter is suspicious since 'Beer, lager, alc. 4.4% by vol.' is listed in the Danish databank as containing 0.05mg Phe and 0.003g protein. It is natural to expect that food containing aspartame will fall outside of the range 20-65 Phe:protein ratio. However, we find 33 foods items such as carrots (65.591 ± 1.455) and peppers (106.977 ± 3.816), which do not contain aspartame. Note that, in the Danish databank, raw carrots (25.714 ± 9.898) and sweet, green, raw peppers (30 ± 8.889) are included in the range 20-65.

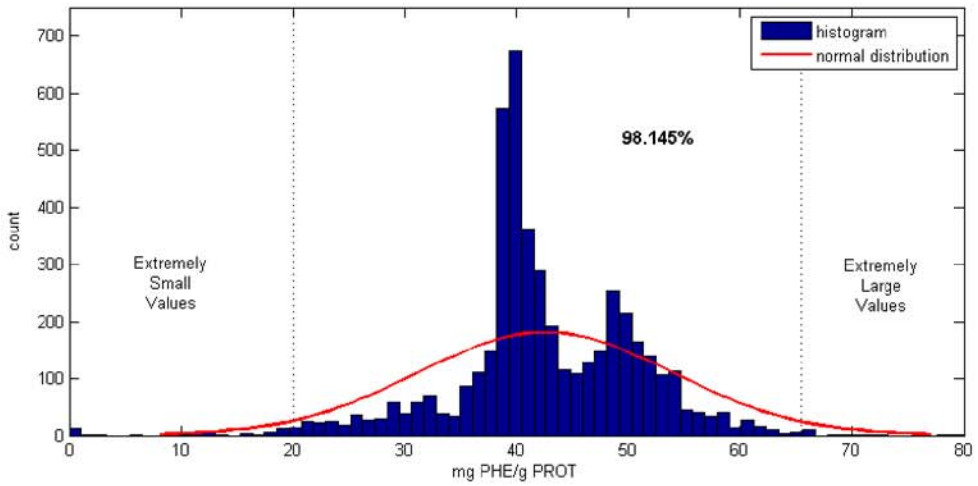


Figure 2.1.: Distribution of Phe:protein ratio for all foods in the USDA database

2.2.2 Danish databank

Figure 2.2 describes the distribution of Phe:protein ratio for all food in the Danish databank. The corresponding normal approximation of the data is plotted with empirical mean (41.373) and standard deviation (15.643). The empirical mean is consistent

with that of the USDA database but not the standard deviation (See Table 2.1). The errors of the Phe:protein ratios in the Danish databank are < 10 for 84.046% and < 5 for 63.960% of the cases. There exists 2 food items with Phe:protein ratios above 90, canned pumpkin (242.5 ± 31.562) and raw pumpkin (250 ± 21.667). These values differ widely from the data in the USDA database, canned pumpkin (31.818 ± 1.988) and raw pumpkin (32 ± 0.66). In the Danish databank, the multipliers 30-50 apply to 71.083%, of all items. On the other hand, $20 : 1 \leq$ Phe:protein $\leq 65 : 1$ is true for 90.122% of listed foods. Hence 20 and 65, once again, are significantly more accurate multipliers. The Phe:protein ratio in 'Fruit and fruit products' has empirical mean 34.528 and standard deviation 14.964. The Phe:protein ratio in 'Vegetables and vegetable products' has empirical mean 37.459 and standard deviation 28.558. Again, while the means are similar to those for the USDA database, the standard deviation is much higher.

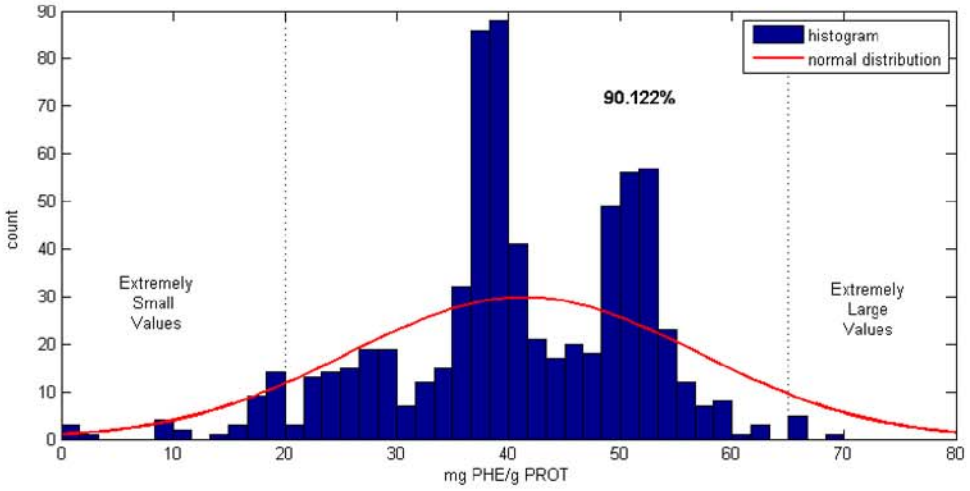


Figure 2.2.: Distribution of Phe:protein ratio for all foods in the Danish databank

Table 2.1.: List of empirical mean and standard deviation for types of food from two sources, USDA database (US), Danish databank (D) and the combined results from two databases (C). A more complete list of Phe:protein ratios can be found in [12].

Category	**Src**	**Average**	**Standard Deviation**
All Foods	US	42.580	11.469
	D	41.373	15.643
	C	42.425	12.095
Fruits & Fruit Juices	US	30.420	12.558
(Fruits & Fruit Products)	D	34.528	14.964
	C	31.811	13.561
Vegetables & Veg Products	US	39.522	11.432
	D	37.459	28.558
	C	39.163	15.834

2.3 Conclusions

Based on our data, we reject Hypothesis 1 and we propose the multipliers 20 and 65 as a replacement. Indeed multiplying the protein content in grams by 20 and 65 yields accurate (in more than 97% of cases) minimum and maximum bounds, respectively, for the Phe content in mg. We accept Hypotheses 2 and 3, namely that the average multipliers for fruits and vegetables are about 30 and 40, respectively. However, we note that our empirical average for fruits is actually slightly higher (31.811) and that for vegetables is slightly lower (39.163). The multipliers 20 and 65 provide a general estimate for the Phe content of foods based on the protein content. The ingredient list and the Nutrition Facts Label provide information that can be used to refine this initial rough estimate. Methods for doing so will be investigated in Chapter 4.

3. Two simple guidelines to limit the phenylalanine intake from sweets with gelatin

It is critical for individuals with PKU to manage Phe intake daily in order to maintain a certain Phe level. However, the Phe content for a food is not always readily available. Without the Phe information, the basis for the decision about how much to have will be vague. In this chapter, we propose a simple method for estimating a proper amount of intake for a certain type of foods in the case where we do not have any access to the Phe data. The work in this chapter is submitted for publication [13].

3.1 Methods

The main objective of this chapter is sweets (or any other food) whose only source of Phe is gelatin. In other words, the foods we are concerned about contain gelatin, which is a significant source of Phe. However, we assume that no other ingredient used to prepare the food contains any Phe (for example, sugar, color, and flavors) (See Table 1 in [12] for a list of Phe free ingredients). We present two lemmas that yield a food amount that is guaranteed not to contain more than 20 mg of Phe. The first lemma uses only the (ranked) ingredient list and is applicable when no part of any ingredient is removed in the preparation process. The mathematical operation needed to apply this lemma is a simple counting procedure (i.e. counting the rank of gelatin in the list) and thus does not require a calculator. The second lemma uses the serving size and the (rounded) protein content obtained from the Nutrition Facts Label and allows for some parts of ingredients to be removed during the preparation process. The mathematical operation needed to apply this lemma is a single (potentially two digits) division and thus requires a calculator.

Lemma 1 *Given is a food whose only ingredient containing Phe is gelatin. Assume that all ingredients used to prepare the food remain entirely in the food after the preparation process is complete. Let k be the rank of gelatin in the ingredient list. Then, k grams of the food contain less then 20 milligrams of Phe.*

Proof Take y grams of the food. Let $A_1, A_2, \cdots, A_n$ be the amounts of each respective ingredient contained in y grams of the food. We have $\Sigma_1^n A_i = y$. Without loss of generality, we can assume that $A_1 \geq A_2 \geq \cdots \geq A_n$ and that A_k is the amount of gelatin contained in y grams of the food. We have

$$kA_k \leq \sum_{i=1}^{k} A_i \leq \sum_{i=1}^{n} A_i = y.$$

Dividing by k, we obtain

$$A_k \leq \frac{y}{k}$$

Since gelatin contains 17.37 milligrams of Phe per gram [7], the amount of Phe in y grams of the food is

$$\text{Phe in food} = 17.37A_k \leq 17.37\frac{y}{k}.$$

Taking $y = k$, we obtain

$$\text{Phe in food} = 17.37A_k \leq 17.37\frac{k}{k} < 20.$$

■

Lemma 2 *Given is a food whose only ingredient containing Phe is gelatin. Let x be the size of a serving in grams, and let p be the protein content, rounded to the nearest 2δ grams. (Note: In the United States, the protein content is usually rounded to the nearest gram, so $\delta = 0.5$). Then, $\frac{x}{p+\delta}$ grams of the food contains less than 20 milligrams of Phe.*

Proof Let y be the unknown number of grams of the food to be consumed. Since the protein content is rounded, the upper bound for the protein contained in x grams of food is $p + \delta$. Then, an upper bound on the protein content for 1 gram of the food

is $\frac{p+\delta}{x}$. This gives an upper bound for the protein content of y grams of the food as $\frac{y(p+\delta)}{x}$. Since gelatin has 20.292 ± 0.007 milligrams for Phe per gram of protein [12] , one can ingest at most $\frac{y(p+\delta)}{x} \times 20.299$ milligrams of Phe from y grams of the food. Since we want the amount of Phe to be less than 20 ($\frac{y(p+\delta)}{x} \times 20.299 < 20$), we conclude that

$$\frac{y(p+\delta)}{x} < \frac{20}{20.299}.$$

Since $\frac{20}{20.299} < 1$, we have $\frac{y(p+\delta)}{x} < 1$, and so

$$y < \frac{x}{p+\delta}.$$

■

or three digits division and thus may require a calculator.

Both Lemma 1 and Lemma 2 provide upper bounds for the amount of food that would contain less than 20 mg Phe. These two bounds may be different, and neither of these is guaranteed to be tight. However, they provide a quick and simple guideline to help individuals with PKU to manage their diet. Suppose that gelatin is the k^{th} ingredient in the ingredient list of a sweet whose serving size is x and protein content is x with δ rounding error. Then, any value less than $max(\, k,\, \frac{x}{p+\delta}\,)$ is suitable amount of a food for individuals with PKU as long as gelatin is only source of the protein content of the food.

3.2 Results and Discussion

We applied the proposed methods to seven commercial sweets and the results are described in Table 3.1. In the case of ALTOIDS, the results of Lemma 1 and Lemma 2 coincide. However, we see from data that Lemma 2 tends to provide less conservative amount than Lemma 1, as we observe in the examples of Starburst, Brachs, Peeps, Jell-O, Parfait, and GelBites. Nevertheless, the effectiveness of Lemma 1 cannot be neglected as we see from the case of ICE BREAKERS, in which Lemma 1 yields higher acceptable amount (11g) than Lemma 2 (4.6g).

The proposed guidelines suggest proper amounts of foods with gelatin for a person who is allowed to consume up to 20mg of Phe. These guidelines require only simple arithmetic operations and are based on the information provided on the food label. The suggested amounts from these guidelines may be very conservative, as shown in Table 3.1. However, knowing that a couple of bites of a sweet is allowed for an individual with PKU can be reassuring.

Table 3.1.: Upper bounds on food amounts to limit Phe intake to no more than 20mg

Description	Rank of Gelatin	**Lemma 1 (g)**	Serving size (g)	Protein Content (mg)	δ	**Lemma 2 (g)**
ALTOIDS[1]	4	4 (6 pieces)	2	0	0.5	4 (6 pieces)
ICE BREAKERS[2]	11	11 (4.78 pieces)	2.3	0	0.5	4.6 (2 pieces)
Starburst[3]	7	7 (1.4 pieces)	40	0	0.5	80 (16 pieces)
Brachs[4]	6	6 (2.93 pieces)	39	0	0.5	78 (38 pieces)
Peeps	3	3 (0.36 Chicks)	42	1	0.5	28 (3.33 Chicks)
Jell-O[5]	4	4	96	1	0.5	64
Parfait[6]	3	3	110	2	0.5	44
GelBites[7]	2	2	82	3	0.5	23.43

Notes

[1]ALTOIDS peppermint

[2]ICE BREAKERS ICE CUBES Peppermint Gum

[3]Starburst Fruit Chews

[4]Brachs Pastel Candy Corn

[5]Jell-O Strawberry Gelatin Snack

[6]Kroger Strawberry Parfait Naturally & artificially Flavored

[7]Kroger GelBites Strawberry Naturally & Artificially Flavored Gelatin Cubes

4. A method for estimating the nutrient content of commercial foods from their label

Some medical diets require keeping track of one's intake of certain nutrients. In order to do this, individuals need to have access to the nutritional information for food they consume. While many nutrients are listed on the Nutrition Facts Label of commercial foods, the information provided is not complete. Indeed, not all nutrients are listed on the label, and the content for the ones that are listed is rounded. Being able to automatically determine the amount of a nutrient contained in the food, or being able to increase the precision of an amount listed in the Nutrition Facts Label, would thus be helpful.

Unfortunately, the Phe content of commercial foods is not listed on the Nutrition Facts Label, and so individuals with PKU must obtain the Phe information from a food list (e.g., [7, 8, 12]). As these databases only list a limited number of foods, alternative methods for finding the Phe content of foods would be desirable.

In this chapter, we propose to estimate the content of a given nutrient such as Phe by obtaining a minimum bound and a maximum bound for the nutrient amount contained in the food. To do this, we use the food label (Nutrition Fact Label and ingredient list), along with the USDA Standard Reference Database (USDA database [7]).

From the food label, we get the serving size x and the n ingredients used in the recipe. Let A_i denote the weight (in grams) of ingredient i, for $i = 1, \ldots, n$. Since the ingredients are listed in decreasing order of weight, we have $A_i \geq A_{i+1}$. If no part of any ingredient is removed in the preparation process, we thus have

$$x \geq A_1 \geq A_2 \geq \ldots \geq A_n > 0, \tag{4.1}$$

$$A_1 + A_2 + \ldots + A_n = x. \tag{4.2}$$

The food label gives us the rounded up content y^{nut} of many nutrients. Let Δ^{nut} be the rounding error for the content of nutrient "nut". We can look for the amount y_i^{nut} of nutrient "nut" in one gram of ingredient i in the USDA database. If no part of any ingredient is removed in the preparation process, we have

$$y^{nut} - \Delta^{nut} \leq \sum_{i=1}^{n} y_i^{nut} A_i \leq y^{nut} + \Delta^{nut}. \tag{4.3}$$

Bounds for the unknowns A_i can be found using linear programming methods for the optimization problem defined by constraints (4.1)-(4.3). Unfortunately, many commercial foods include ingredients that are not listed in the USDA database: y_i^{nut} is unknown for these ingredients. The optimization problem defined by (4.1)-(4.3) then becomes non-linear.

We propose an alternative method for finding bounds, $A_{i_{min}}$ and $A_{i_{max}}$, for each ingredient amount A_i. The method is iterative, and is applicable even if the nutrient data for some of the ingredients is missing. The bounds obtained this way yield a first set of bounds for the amount of the considered nutrient (e.g., Phe) contained in the food. This is Step 1 of our proposed method for nutrient content estimation, which we describe in Section 4.1. This step requires prior knowledge of the amount of the considered nutrient (e.g., Phe) for each ingredient. For example, when trying to estimate the Phe content of the food, then the Phe contents for all the ingredients must be known. Since many ingredients not listed in the USDA database clearly do not contain a significant amount of proteins (e.g., food coloring, natural flavor, etc.) and thus can be considered free of Phe, this is a reasonable assumption.

In Step 2 of our method, we make use of the Simplex algorithm in order to further narrow the interval of bounds for the nutrient content. This step is described in Section 4.2. Our method (Step 1 and Step 2) is applied to the problem of approximating the ingredient amounts and estimating the Phe content of various commercial foods in Section 4.3. We conclude in Section 4.4. Note that a preliminary version of this work, which contained only Step 1 of our method, was previously presented in a conference paper [14], and this work is submitted for publication [15].

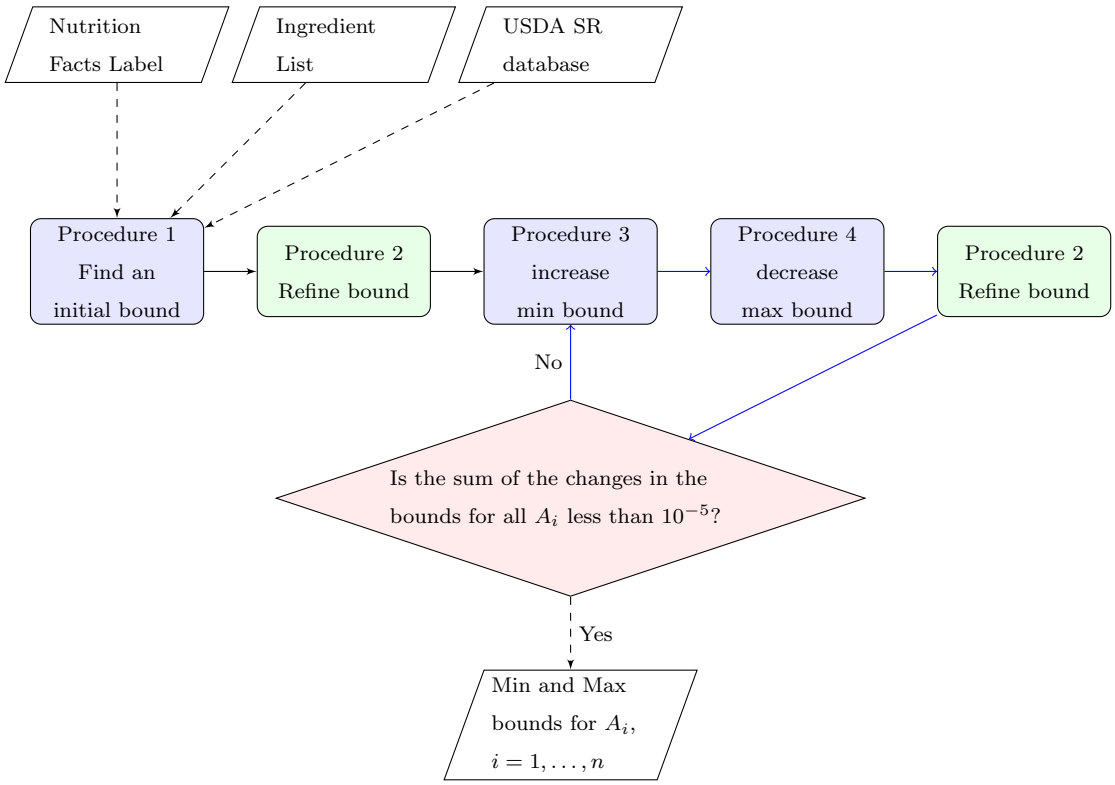


Figure 4.1.: Schematic Diagram of Proposed Method to Estimate the Ingredient Amounts

4.1 Step 1: Nutrient content estimation using Approximate ingredient amounts

If we knew A_i, the amount of ingredient i, along with p_i, the number of milligrams of a given nutrient per gram of ingredient i, then $p_i A_i$ would be the nutrient contributed by ingredient i, and the total given nutrient in the food would be $\Sigma_{i=1}^{n} p_i A_i$. Therefore, we have the following bounds for the nutrient content (NUT),

$$\Sigma_{i=1}^{n} p_i A_{i_{min}} \leq NUT \leq \Sigma_{i=1}^{n} p_i A_{i_{max}}. \tag{4.4}$$

We now propose a simple technique to obtain an initial range estimate ($A_{i_{min}}$, $A_{i_{max}}$) for each ingredient i, along with an iterative method to refine the bounds of the range. These estimates shall then be put into Equation (4.4) to obtain a first set of bounds for the content of the considered nutrient NUT.

4.1.1 Initial range estimate

Our initial range estimates are based on the following two lemmas.

Lemma 3 *If* $\{A_i\}_{i=1}^{n}$ *satisfy Equation (4.1) and (4.2), then*

$$\frac{x}{n} \leq A_1 \leq x, \tag{4.5}$$

$$0 < A_i \leq \frac{x}{i}, \quad for\, i = 2, 3, \ldots, n. \tag{4.6}$$

Proof Since $A_i \leq A_{i-1} \leq \ldots \leq A_1$,

$$iA_i \leq \sum_{k=1}^{i} A_k \leq \sum_{k=1}^{n} A_k = x.$$

Dividing each side by i, we get

$$A_i \leq \frac{x}{i}.$$

In addition,

$$x = \sum_{k=1}^{n} A_k \leq \sum_{k=1}^{n} A_1 = nA_1.$$

Dividing each side by n, we have a minimum bound for A_1,

$$\frac{x}{n} \leq A_1.$$

■

Lemma 4 *If $\{A_i\}_{i=1}^n$ satisfy Inequality (4.3) and $y_i^{nut} \neq 0$, then*

$$A_i \leq \frac{y^{nut}}{y_i^{nut}}, \quad \text{for } i = 1, 2, \ldots, n. \tag{4.7}$$

Equality holds for some i_0 only if ingredient i_0 is the sole ingredient containing the nutrient.

Proof Suppose that

$$A_i > \frac{y^{nut}}{y_i^{nut}}, \quad \text{for some } i.$$

This implies $y_i^{nut} A_i > y^{nut}$. However,

$$y_i^{nut} A_i \leq \sum y_j^{nut} A_j = y^{nut}, \quad \text{for all } i.$$

This is a contradiction, so

$$A_i \leq \frac{y^{nut}}{y_i^{nut}}$$

is true for all i. Now considering equality on (4.7), assume that there are more than one ingredient containing the nutrient. If $A_{i_0} = \frac{y^{nut}}{y_{i_0}^{nut}}$, then $y_{i_0}^{nut} A_{i_0} = y^{nut}$. Therefore,

$$\begin{aligned} y^{nut} &= \textstyle\sum_{i=1}^n y_i^{nut} A_i \\ &= y_{i_0}^{nut} A_{i_0} + \textstyle\sum_{i=1, i\neq i_0}^n y_i^{nut} A_i \\ &= y^{nut} + \textstyle\sum_{i=1, i\neq i_0}^n y_i^{nut} A_i \end{aligned}$$

This gives

$$0 = \sum_{i=1, i\neq i_0}^{n} y_i^{nut} A_i.$$

But,

$$\sum_{i=1, i\neq i_0}^{n} y_i^{nut} A_i > 0$$

since there exists another index t then i_0 such that $y_t^{nut} > 0$ by assumption. This is a contradiction. ■

The initial bounds for each A_i are obtained by combining Equation (4.5), (4.6) and (4.7), as described in Procedure 1. Note that the Procedure takes into account the rounding error (Δ^{nut}) in the nutrient contents listed on the food label and the rounding errors (Δ_i^{nut}) in the USDA database.

Procedure 1 Initial bound

$A_{1_{min}} \leftarrow \frac{x}{n}, \quad A_{1_{max}} \leftarrow x$

for $i = 2$ to n **do**

$A_{i_{min}} \leftarrow 0, \quad A_{i_{max}} \leftarrow \frac{x}{i}$

end for

for given nutrient with content y^{nut} **do**

if $y_1^{nut} \neq 0$ **then**

$A_{1_{max}} \leftarrow min(A_{1_{max}}, \frac{y^{nut}+\Delta^{nut}}{y_1^{nut}-\Delta_1^{nut}})$

end if

for $i = 2$ to n **do**

if $y_i \neq 0$ **then**

$A_{i_{max}} \leftarrow min(A_{i_{max}}, A_{i-1_{max}}, \frac{y^{nut}+\Delta^{nut}}{y_i^{nut}-\Delta_i^{nut}})$

else

$A_{i_{max}} \leftarrow min(A_{i_{max}}, A_{i-1_{max}})$

end if

end for

end for

4.1.2 Iterative method to narrow the range estimate

The initial bounds $A_{i_{min}} \leq A_i \leq A_{i_{max}}$ can be refined using the equation $x = \Sigma_{i=1}^{n} A_i$. More specifically, we have

$$x - \Sigma_{j=1,j\neq i}^{n} A_{j_{max}} \leq A_i \leq x - \Sigma_{j=1,j\neq i}^{n} A_{j_{min}}$$

and so Procedure 2 can be used to narrow the range of each A_i.

Procedure 2 Refining bound

for $i = 1$ to n **do**

$A_{i_{min}} \leftarrow max(A_{i_{min}}, x - \Sigma_{j=1,j\neq i}^{n} A_{j_{max}})$

$A_{i_{max}} \leftarrow min(A_{i_{max}}, x - \Sigma_{j=1,j\neq i}^{n} A_{j_{min}})$

end for

More refinement can be obtained using the following lemma.

Lemma 5 *Suppose $A_{i_{min}} \leq A_i \leq A_{i_{max}}$ for $i = 1, \ldots, n$. If $y_k^{nut} \neq 0$ for some k, then*

$$A_k \leq \frac{y^{nut} - y_i^{nut} A_{i_{min}}}{y_k^{nut}} \quad \text{for all } i \neq k.$$

Also,

$$A_k \leq \frac{y^{nut} - \Sigma_{i=1,i\neq k}^{n} y_i^{nut} A_{i_{min}}}{y_k^{nut}}.$$

Furthermore, if y_i^{nut} is known for all i,

$$A_k \geq \frac{y^{nut} - \Sigma_{i=1,i\neq k}^{n} y_i^{nut} A_{i_{max}}}{y_k^{nut}}.$$

Proof Since $\sum_{i=1}^{n} y_i^{nut} A_i = y^{nut}$, we have

$$y_k^{nut} A_k = y^{nut} - \sum_{i=1,i\neq k}^{n} y_i^{nut} A_i \tag{4.8}$$

Multiplying y_i^{nut} to the set of initial bound for A_i, we get the set of bound for $y_i^{nut} A_i$ such that

$$y_i^{nut} A_{i_{min}} \leq y_i^{nut} A_i \leq y_i^{nut} A_{i_{max}} \quad \text{for all } i.$$

Then by (4.8),

$$\begin{aligned} y_k^{nut} A_k = &\quad y^{nut} - \textstyle\sum_{j=1, j\neq k}^{n} y_j^{nut} A_j \\ \leq &\quad y^{nut} - y_i^{nut} A_i, \qquad \text{for all } i \neq k, \\ \leq &\quad y^{nut} - y_i^{nut} A_{i_{min}}, \qquad \text{for all } i \neq k. \end{aligned}$$

Since $y_k^{nut} > 0$, dividing each side by y_k^{nut} yields

$$A_k \leq \frac{y^{nut} - y_i^{nut} A_{i_{min}}}{y_k^{nut}}.$$

Furthermore, if y_i^{nut} is known for all i,

$$\sum_{i=1, i\neq k}^{n} y_i^{nut} A_{i_{min}} \leq \sum_{i=1, i\neq k}^{n} y_i^{nut} A_i \leq \sum_{i=1, i\neq k}^{n} y_i^{nut} A_{i_{max}}.$$

Thus,

$$y^{nut} - \sum_{i=1, i\neq k}^{n} y_i^{nut} A_{i_{max}} \leq y^{nut} - \sum_{i=1, i\neq k}^{n} y_i^{nut} A_i,$$

and

$$y^{nut} - \sum_{i=1, i\neq k}^{n} y_i^{nut} A_i \leq y^{nut} - \sum_{i=1, i\neq k}^{n} y_i^{nut} A_{i_{min}}.$$

Combining these inequalities with Equation (4.8),

$$y^{nut} - \sum_{i=1, i\neq k}^{n} y_i^{nut} A_{i_{max}} \leq y_k^{nut} A_k \leq y^{nut} - \sum_{i=1, i\neq k}^{n} y_i^{nut} A_{i_{min}}.$$

Therefore, dividing each side by y_k^{nut}, we can conclude that

$$\frac{y^{nut} - \sum_{i=1, i\neq k}^{n} y_i^{nut} A_{i_{max}}}{y_k^{nut}} \leq A_k$$

and

$$A_k \leq \frac{y^{nut} - \sum_{i=1, i\neq k}^{n} y_i^{nut} A_{i_{min}}}{y_k^{nut}}. \tag{4.9}$$

Even though y_i^{nut} is unknown for some i, the maximal bound for A_k given in Equation (4.9) is still reasonable if we set zero for the unknown y_i^{nut}. ■

Procedure 3 to increase the minimal bound

for given nutrient with content y^{nut} such that y_k^{nut} exists $\forall k$ **do**

if $y_n^{nut} \neq 0$ **then**

$A_{n_{min}} \leftarrow max(A_{n_{min}}, \frac{(y^{nut}-\Delta^{nut})-\Sigma_{k=1}^{n-1}(y_k^{nut}+\Delta_k^{nut})A_{k_{max}}}{y_n^{nut}+\Delta_n^{nut}})$

end if

for $i = n-1$ to 1 **do**

if $y_i^{nut} \neq 0$ **then**

$A_{i_{min}} \leftarrow max(A_{i_{min}}, A_{i+1_{min}}, \frac{(y^{nut}-\Delta^{nut})-\Sigma_{k=1,k\neq i}^{n}(y_k^{nut}+\Delta_k^{nut})A_{k_{max}}}{y_i^{nut}+\Delta_i^{nut}})$

end if

end for

end for

Procedure 4 to decrease the maximal bound

for given nutrient with content y^{nut} **do**

for $k = 1$ to n **do**

if y_k^{nut} does not exist **then**

$y_k^{nut} \leftarrow 0, \quad \Delta_k^{nut} \leftarrow 0$

end if

end for

if $y_1^{nut} \neq 0$ **then**

$A_{1_{max}} \leftarrow min(A_{1_{max}}, \frac{(y^{nut}+\Delta^{nut})-\Sigma_{k=2}^{n}(y_k^{nut}-\Delta_k^{nut})A_{k_{min}}}{y_1^{nut}-\Delta_1^{nut}})$

end if

for $i = 2$ to n **do**

if $y_i^{nut} \neq 0$ **then**

$A_{i_{max}} \leftarrow min(A_{i_{max}}, A_{i-1_{max}}, \frac{(y^{nut}+\Delta^{nut})-\Sigma_{k=1,k\neq i}^{n}(y_k^{nut}-\Delta_k^{nut})A_{k_{min}}}{y_i^{nut}-\Delta_i^{nut}})$

end if

end for

end for

Lemma 5 yields methods to increase the minimal bound (Procedure 3) and to decrease the upper bound (Procedure 4). Note that the minimal bound can only be refined if y_i^{nut} is known for all i. Otherwise, the bound remains as it is. This is not the case for the maximal bound.

Let us summarize our proposed Step 1. To estimate the A_i's, we first select a set of nutrients that are listed on the Nutrition Facts Label (e.g., carbohydrates, sodium, protein, etc.). We then apply Procedure 1 (running over all selected nutrients), followed by Procedure 2. After that, we keep repeating Procedure 3 and Procedure 4 (running over all selected nutrients), followed by Procedure 2, until our estimates change by less than 10^{-5} between consecutive repetitions. This is illustrated in Figure 4.1.

4.2 Step 2: Nutrient content estimate Refinement using Simplex algorithm

Observe that the bounds obtained using Equation (4.4) correspond to ingredient amounts that can violate Equation (4.2). More specifically, neither $\sum_{i=1}^{n} A_{i_{min}}$ nor $\sum_{i=1}^{n} A_{i_{max}}$ equal to a serving size x in general. This indicates that it should be possible to further refine the content estimate obtained in Step 1. We propose to do this using the Simplex algorithm [16] which is a well-known linear programming tool. The Simplex algorithm first finds an initial feasible solution in Phase I. Then, in Phase II, it moves along the edges of the polytope defined by the constraints while evaluating the cost until it reaches an extreme value. In the case of the nutrient content estimation problem, the cost function is the summation of the nutrient content coming from each ingredient, the nutrient content (NUT).

$$\text{cost} = \sum_{i=1}^{n} p_i A_i \quad (= NUT).$$

There are different ways to write the linear constraints of the problem. We start from the constraints obtained in Step 1: $A_{i_{min}} \leq A_i \leq A_{i_{max}}$. We then introduce new nonnegative variables,

$$a_n = A_n,$$

$$a_i = A_i - A_{i+1}, \qquad \text{for} \quad i = 1, \ldots, n-1,$$

$$d_i \leq A_{i_{max}} - A_{i_{min}}, \quad \text{for} \quad i = 1, 2, \ldots, n,$$

and slack variables $s_i \geq 0$ for $i = 1, 2, \ldots, n$ such that

$$A_i + d_i = A_{i_{max}},$$

$$d_i + s_i = A_{i_{max}} - A_{i_{min}}.$$

Then the amount of ingredient i is given by the summation of a_k for $k = i, \ldots, n$,

$$A_i = A_n + \sum_{k=i}^{n-1}(A_k - A_{k+1}) = \sum_{k=i}^{n} a_k.$$

We can also rewrite Equation (4.2) in terms of these new variables,

$$x = \sum_{i=1}^{n} A_i = \sum_{i=1}^{n}\sum_{k=i}^{n} a_k = \sum_{i=1}^{n} i a_i.$$

Secondly, the nutrient content (NUT) can be obtained by

$$NUT = \sum_{i=1}^{n} p_i A_i = \sum_{i=1}^{n} p_i \sum_{k=i}^{n} a_k = \sum_{i=1}^{n}(\sum_{m=1}^{i} p_m) a_i.$$

Lastly, by subtracting $\sum_{k=i+1}^{n} a_k + d_{i+1} = A_{i+1_{max}}$ from $\sum_{k=i}^{n} a_k + d_i = A_{i_{max}}$, we have the constraints

$$a_i + d_i - d_{i+1} = A_{i_{max}} - A_{i+1_{max}}$$

for $i = 1, 2, \ldots, n-1$. Therefore, we define the nutrient content estimation problem by Definition 1.

Since all constraints are equalities, any feasible solutions satisfying the constraints are points on the edges of a (n-1)-dimensional polytope. Hence, once an initial feasible

point is found from Phase I of the Simplex algorithm, in Phase II, we look through the extreme points of the polytope until the cost at any adjacent points of an extreme point does not decrease anymore. The cost at the point becomes the minimum of the nutrient content for a serving size x gram of a food. Similarly, once the cost function does not increase anymore, we set the maximum bound for the nutrient content to the value of the cost function.

Definition 1 Nutrient content estimate using Simplex algorithm

minimize, maximize $\sum_{i=1}^{n}(\sum_{k=1}^{i} p_k)a_i$ *where*

$$\begin{cases} \sum_{i=1}^{n} ia_i = x, & \\ a_i + d_i - d_{i+1} = A_{i_{max}} - A_{i+1_{max}}, & i = 1, \ldots, n-1, \\ a_n + d_n = A_{n_{max}}, & \\ d_i + s_i = A_{i_{max}} - A_{i_{min}}, & i = 1, \ldots, n, \\ a_i, d_i, s_i \geq 0, & i = 1, \ldots, n. \end{cases}$$

4.3 Numerical Experiments

4.3.1 Convergence of ingredient amounts (A_i)

To directly test our method for estimating the amount of each ingredient in a commercial food, we would need to have the true ingredient amounts. For good reasons, manufacturers are unwilling to share this information. However, we can test the accuracy of our method by looking at the difference between the estimated maximum and the estimated minimum. The difference should become smaller as we consider more nutrients, indicating convergence to the true values.

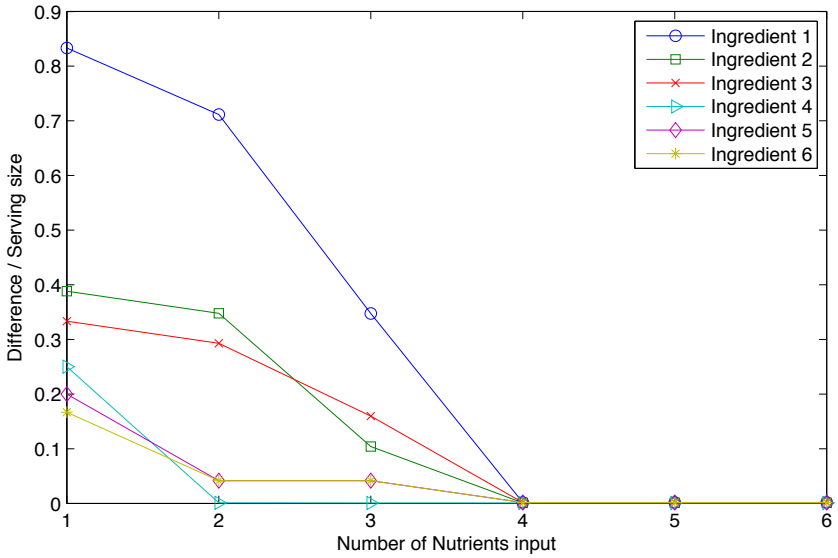
To test this, we estimated the ingredient amounts of various foods using a subset (in order) of the following nutrients: protein, sodium, energy, carbohydrates, fat, and cholesterol. Some of our results are illustrated in Figure 4.2. As expected, the

estimated maximum/minimum ingredient amounts tended to decrease/increase as we considered more nutrients. As a result, the range of the estimates (measured by the ratio of the difference between the maximum and the minimum and the serving size in Figure 4.2) decreased. In some cases (e.g., Spicy Brown Mustard in Figure 4.2(a)), the range decreased to nearly zero ($< 0.2\,\%$ serving size) for all ingredients with only 4 nutrients. In other cases (e.g., Garlic mashed potatoes in Figure 4.2(b)), we failed to obtain a good estimates for some of the ingredients even though we obtained a near perfect estimate for the other ingredients with just two nutrients. Clearly, the accuracy of our method depends on the food considered and can vary for one ingredient to the next. However it is not necessary that all ingredient amounts be precisely estimated in order to get a good estimate on the content of the query nutrient NUT, as we shall see in the following.

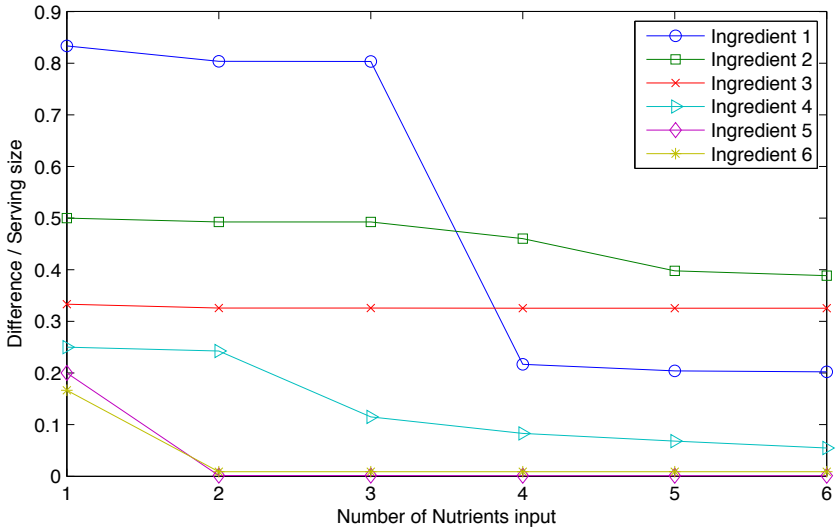
4.3.2 Application to Phenylalanine (Phe) content estimation

We experimented with our method to estimate the Phenylalanine (Phe) content for 25 commercial food items. The results are shown in Table 4.1 and 4.2. They present the minimum and maximum bounds for each food obtained by our method using six nutrients (protein, sodium, energy, carbohydrates, fat, and cholesterol). Both the results after Step 1 (column 4) and Step 2 (column 5) are given in order to see the improvement resulting from performing Step 2.

For comparison, the Phe data from two databases, USDA database [7] and a low-protein food database [8], are written in the first and second column of Table 4.1 and 4.2. When there exists no data related to the item from that database, we indicated the case with 'N/A'. As we expected, only a part of the food items considered has Phe data in the USDA database (6/25) or Phe data in the low-protein food database (14/25). Furthermore, some of the data listed in our table may be inexact as we were unable to find the specific brand of product considered and used a generic version instead. For example, the Phe content for Tomato soup specifically from Campbell



(a) Spicy brown mustard



(b) Simply potatoes garlic mashed potatoes

Figure 4.2.: **Range of Estimates for Ingredient Amounts.** As more nutrients are taken into account, the difference between the estimated maximum amount and the estimated minimum amount for each ingredient often decreases quickly.

company is not presented in the USDA database while the USDA database contains the Phe content of Tomato soup for any brand. For additional comparison, we performed the full linear optimization defined by the cost function $\sum_{i=1}^{n} p_i A_i$ (where p_i denotes the number of milligrams of Phe per one gram of ingredient i) and constraints (4.1), (4.2), and (4.3). The Phe content was optimized only using the Simplex algorithm, not applying the proposed approximate method. The approximated minimum and maximum bounds are written in parentheses in the third column of Table 4.1 and 4.2. Because of the overdetermined full linear system, the Simplex algorithm failed to find an initial feasible solution for most of the cases (17/25); these are denoted by 'DNE[c]' in the table. We also could not obtain a result from the Simplex algorithm in case of 'Vinaigrette Balsamic Dressing' because of missing nutrient data for one of the ingredients in the USDA database.

In contrast, Step 1 of our approximate method was able to provide bounds for the Phe content of all targeted food items, as shown in the fourth column of Table 4.1 and 4.2 where the estimated minimum and maximum values for the Phe content are written in parenthesis. The range between the minimum and the maximum bounds was less than 10mg for 16 food items, and less than 25mg for 19 food items. The estimated bounds for the Phe content were within no more than 3 mg from at least one of the databases for 22 items, which is 88% of the 25 foods considered. In the case of butter, rice krispies cereal and waffles, our range excluded the Phe value from both databases. This is most likely due to the violation of our assumption that no part of any ingredient is discarded during the preparation process. For example, there is considerable drying in the preparation of cereal, and liquid (whey) is discarded in the preparation of butter.

After Step 2, the interval between bounds for the Phe content narrowed significantly more in 10 cases (see the fifth column of Table 4.1 and 4.2). Step 2 narrowed the range of the estimated Phe bounds for one serving of Salsa Sauce from 24.68mg to 10.33mg. In the case of garlic mashed potatoes and sweet potato tot, the ranges

Table 4.1.: Comparison of phenylalanine content estimates obtained with our methods, two food databases and the full linear programming approach (Simplex Algorithm) (1).

Description (serving size)	**USDA database [7]**	**low-protein food database [8]**	**Full linear optimization**	**After Step 1**	**After Step 2**
Carr's Whole Wheat Crackers (17 g)	81.6 mg	75 mg	(57.68 mg, 78.35 mg)	(53.61 mg, 85.11 mg)	(53.61 mg, 85.11 mg)
Ketchup (17 g)	4.42 mg	10.2 mg	(1.44 mg, 4.42 mg)	(0.70 mg, 7.09 mg)	(1.20 mg, 6.57 mg)
KIT KAT Milk Chocolate (42 g)	113.4 mg	131.86 mg	DNE[8]	(129.56 mg, 238.91 mg)	(144.27 mg, 191.53 mg)
Campbell's Tomato soup (122 g)	68.32 mg[9]	66.90 mg	DNE[8]	(33.21 mg, 102.91 mg)	(40.69 mg, 95.45 mg)
Cheerios Cereal (28 g)	175.84 mg	165 mg	DNE[8]	(179.86 mg, 180.51 mg)	DNE[8]
Rice Krispies Cereal (33 g)	116.82 mg	107 mg	DNE[8]	(91.54 mg, 94.80 mg)	DNE[8]
Enchilada Sauce (60 g)	N/A	6 mg	(1.53 mg, 24.83 mg)	(0.41 mg, 35.69 mg)	(0.41 mg, 34.14 mg)
Eggo waffle (70 g)	N/A	238 mg	(196.73 mg, 216.09 mg)	(196.26 mg, 216.35 mg)	(196.26 mg, 216.35 mg)
Garlic chili pepper sauce (9 g)	N/A	1.93 mg	(2.71 mg, 5.27 mg)	(1.37 mg, 6.96 mg)	(2.65 mg, 5.27 mg)
Salsa sauce (30 g)	N/A	11 mg	(9.12 mg, 18.20 mg)	(1.53 mg, 26.21 mg)	(7.90 mg, 18.23 mg)
Garlic mashed potatoes (124 g)	N/A	N/A[10]	(154.71 mg, 158.29 mg)	(56.89 mg, 222.50 mg)	(139.51 mg, 162.23 mg)
Butter with Canola Oil (14 g)	N/A	6 mg	DNE[8]	(11.88 mg, 17.66 mg)	(12.06 mg, 17.66 mg)
Go-Gurt (64 g)	N/A	120 mg	DNE[8]	(116.38 mg, 120.95 mg)	DNE[8]
Jell-O Gelatin Snacks (98 g)	N/A	23.76 mg	DNE[8]	(10.01 mg, 30.44 mg)	(10.01 mg, 30.44 mg)
Marshmallow Peeps, Baby Chicks (42 g)	N/A	21 mg	DNE[8]	(19.17 mg, 23.56 mg)	DNE[8]
Ore-Ida French fries (84 g)	N/A	76 mg	DNE[8]	(77.64 mg, 78.77 mg)	(77.64 mg, 78.76 mg)
Spicy Brown Mustard (5 g)	N/A	8 mg	DNE[8]	(9.87 mg, 10.35 mg)	(10.11 mg, 10.16 mg)
Starburst Fruit Chews (40 g)	N/A	5.42 mg	DNE[8]	(0.00 mg, 4.48 mg)	DNE[8]
Vinaigrette Balsamic Dressing (31 g)	N/A	3 mg	DNE[11]	(0.00 mg, 5.53 mg)	(0.00 mg, 5.53 mg)
Yoplait Original Strawberry (170 g)	N/A	284.67 mg	DNE[8]	(287.11 mg, 291.08 mg)	DNE[8]

Table 4.2.: Comparison of phenylalanine content estimates obtained with our methods, two food databases and the full linear programming approach (Simplex Algorithm) (2).

Description (serving size)	**USDA database [7]**	**low-protein food database [8]**	**Full linear optimization**	**After Step 1**	**After Step 2**
ALTOIDS peppermint (2 g)	N/A	N/A	DNE[8]	(0.43 mg, 4.22 mg)	DNE[8]
Jell-O Cheesecake Pudding Dessert (26 g)	N/A	N/A	DNE[8]	(0.91 mg, 0.98 mg)	DNE[8]
Sweet potato Tot (85 g)	N/A	N/A	DNE[8]	(54.87 mg, 113.77 mg)	(71.91 mg, 95.82 mg)
Taco Shells (32 g)	N/A	N/A	DNE[8]	(36.69 mg, 38.31 mg)	(36.69 mg, 38.31 mg)
Vanilla bean Ice cream (87 g)	N/A	N/A	DNE[8]	(206.87 mg, 211.09 mg)	DNE[8]

Notes

[8]Simplex algorithm could not find a solution

[9]Any brand Tomato soup, condensed. Not Campbell's product.

[10]Database has a value, but with a different protein content.

[11]Simplex algorithm is not applicable due to missing data.

of the estimated bounds for Phe content decreased to the values less than half of the ranges after Step 1. Moreover, the largest range between the minimum and maximum bounds after Step 2 became 54.76mg, less than one third of the highest range after Step 1 (165.61mg). The Simplex algorithm in Step 2 could not find an initial feasible solution for 9 items. This could be because an ingredient used to prepare the food did not coincide with the ingredient listed in the USDA database. Another inconsistency could have occurred because we neglected ingredients with negligible amounts for which the USDA database did not provide any data. However, even though we could not improve the bounds for the Phe content any further for these 9 items after Step 2, notice that the bounds after Step 1 in these cases were already very close to each other, with a difference of less than 5mg per serving size.

4.4 Summary and Conclusions

The Food Safety and Inspection Service of the USDA (United States Department of Agriculture) mandates food companies to label their products with an ingredient list and a Nutrition Facts Label. This information is important, but incomplete. Indeed, some nutrients such as Phenylalanine (Phe) are not listed on the label. This is problematic for individuals with inherited metabolic disorders such as PKU who must carefully monitor their Phe intake. Thus, we propose a method for estimating the content of a given nutrient automatically from the food label information. The method also produces bounds on the amount of each ingredient used to prepare the food, so it can be used as an approximate inverse recipe method.

We assume that no part of any ingredient is removed while preparing a food. This gives two constraints: the sum of each ingredient content equals to a serving size for the food and the weighted sum of a nutrient content for one gram of each ingredient equals to the nutrient content for one serving of the food. We also use the fact that the ingredients are listed in decreasing amounts (per weight). The proposed method is applicable even if the nutrient content of some of the ingredients is not fully known.

But, in general, the more nutrient information is known, the better the accuracy of the final estimate, as measured by the difference between the final maximum bound and minimum bound on the given nutrient amount.

We applied our method to the problem of Phe content estimation. Our approach finds bounds for the Phe content of a food. Step 1 finds minimum and maximum bounds for the ingredient amounts using an iterative method. A first set of minimum and maximum bounds are then obtained from these ingredient amount bounds. Step 2 refines the results using linear programming (Simplex algorithm). We showed our results for various commercial foods in Table 4.1 and 4.2. The intervals between the estimated bounds for the Phe content after Step 2 were within 10.4mg for 17 items and within 24mg for 21 items out of the 25 foods considered. In contrast, the intervals were within 10mg for 16 items and within 25mg for 19 items after Step 1.

While two current databases did not contain Phe data for all the food we considered, our method provided a Phe content estimate for all of them. Hence, we believe that our work provides a useful tool to help individuals with PKU to manage their diet. Moreover, our method can be used to estimate other nutrient contents, or to increase the precision of the nutrient content listed on the Nutrition Facts Label. So it should be helpful in managing other diets as well.

5. Conclusion

One of the difficulties in keeping track of one's nutritional intake (i.e. when managing a metabolic disease) is the lack of readily available nutritional information. For addressing this issue in the case of commercial foods, we proposed new mathematical reasoning and computational methods to estimate the food nutrient content.

We illustrated this idea in the specific case of phenylalanine to limit the phenylalanine daily intake for the treatment of phenylketonuria. The framework of this research is founded on statistical distribution, properties of inequality, and linear programming. Similar approaches could be derived for other amino acids or nutrients, as well as other categories of foods. We hope that the example we presented demonstrate the effectiveness of using mathematics for food nutrient content estimation, and that this work will trigger interest in this new research topic.

In two of the three approaches we propose to estimate the phenylalanine content, we made an assumption that there is no loss of any ingredient during the preparation process. However, this is not always the case; for example, the water of the cream is discarded when preparing butter. Therefore, future research can be done on generalizing our methods to be applicable without this assumption. In addition, our current web and Android applications demand that users input multiple components of the Nutrition Facts Label and the ingredient list. To improve our applications to be more user-friendly, typing each data can be replaced by optical character recognition (OCR) of food label or barcode. With or without these improvements, we believe that our research will help individuals to more easily control their dietary intakes.

REFERENCES

REFERENCES

[1] Ellen Song Kang, Natalie D Sollee, and Park S Gerald. Results of treatment and termination of the diet in phenylketonuria (pku). *Pediatrics*, 46(6):881, 1970.

[2] National Institutes of Health Consensus Development Panel et al. National institutes of health consensus development conference statement: phenylketonuria: screening and management, october 16–18, 2000. *Pediatrics*, 108(4):972–982, 2001.

[3] I Özalp, T Coşkun, M Ceyhan, S Tokol, O Oran, G Erdem, G Takinalp, Z Durmuş, and Y Tarikahya. Incidence of phenylketonuria and hyperphenylalaninaemia in a sample of the turkish newborn population. In *Practical Developments in Inherited Metabolic Disease: DNA Analysis, Phenylketonuria and Screening for Congenital Adrenal Hyperplasia*, pages 237–239. Springer, 1986.

[4] Derek A Applegarth, Jennifer R Toone, et al. Incidence of inborn errors of metabolism in british columbia, 1969–1996. *Pediatrics*, 105(1):e10–e10, 2000.

[5] Simon Sanderson, Anne Green, MA Preece, and Hilary Burton. The incidence of inherited metabolic disorders in the west midlands, uk. *Archives of disease in childhood*, 91(11):896–899, 2006.

[6] Jerry Vockley, Hans C Andersson, Kevin M Antshel, Nancy E Braverman, Barbara K Burton, Dianne M Frazier, John Mitchell, Wendy E Smith, Barry H Thompson, Susan A Berry, et al. Phenylalanine hydroxylase deficiency: diagnosis and management guideline. *Genetics in Medicine*, 16(2):188–200, 2013.

[7] U.S. Department of Agriculture, Agricultural Research Service. USDA national nutrient database for standard reference, release 25. Nutrient Data Laboratory Home Page, `http://www.ars.usda.gov/ba/bhnrc/ndl`, 2012.

[8] Virginia E. Schuett. *Low Protein Food List for PKU*. Third edition, 2010.

[9] E. Saxholt, A. Christensen, A.T.and Mller, H.B. Hartkopp, K. Hess Ygil, and O.H. Hels. Danish food composition databank, revision 7. Department of Nutrition, National Food Institute, Technical University of Denmark. Available at: `http://www.foodcomp.dk/`, 2008.

[10] Annabel L Sweeney, Rachel Margaret Roberts, and Janice M Fletcher. Dietary protein counting as an alternative way of maintaining metabolic control in phenylketonuria. In *JIMD Reports-Case and Research Reports, 2011/3*, pages 131–139. Springer, 2012.

[11] Jieun Kim and Mireille Boutin. New multipliers for estimating the phenylalanine content of foods from the protein content. *Journal of Food Composition and Analysis*, 2015.

[12] Jieun Kim and Mireille Boutin. A list of phenylalanine to protein ratios for common foods. ECE Technical Reports. Paper 456. Available at: `http://docs.lib.purdue.edu/ecetr/456`, 2014.

[13] Jieun Kim and Mireille Boutin. Deriving nutritional information using mathematics: the example of phenylalanine in sweets with gelatin. submitted, 2015.

[14] Jieun Kim and Mireille Boutin. An approximate inverse recipe method with application to automatic food analysis. In *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on*, pages 32–39. IEEE, 2014.

[15] Jieun Kim and Mireille Boutin. A method for estimating the nutrient content of commercial foods from their label. submitted, 2015.

[16] Mongi Benhamadou. On the simplex algorithm revised form. *Advances in Engineering Software*, 33(11):769–777, 2002.

VITA

VITA

Jieun Kim was born in Seoul, South Korea. She received her B.S. degree in Mathematics from Hanyang University, Seoul, South Korea in 2008. She began pursuing her PhD in the Department of Mathematics at Purdue University, West Lafayette, Indiana in 2009. Meanwhile, she also worked towards an M.S. degree in the School of Electrical and Computer Engineering at the same academic institution. Her research interest includes the development of mathematical and statistical frameworks for the inverse recipe problem and food nutrient content estimation.