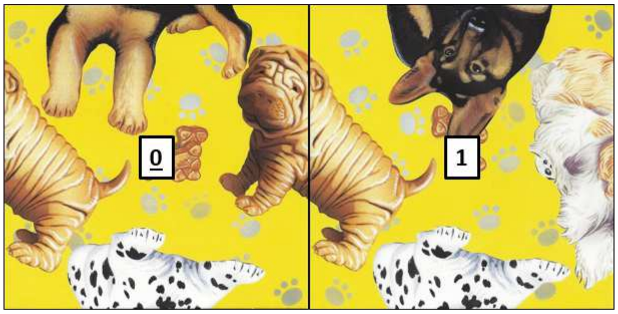
Approaching Protein Folding through Scramble Squares

Scramble Squares is a puzzle belonging to the “tile matching & placement” genre of classical *combinatorial* problems. While this puzzle might seem like just a childhood novelty, there is a surprising degree of similarity between this puzzle and the more compelling challenge of predicting protein folding.

Understanding why it is wholly intractable (impossible) to solve a Scramble Squares puzzle using only a naïve brute force enumeration technique without the aid of any heuristics (rules of thumb), sheds a lot of light as to why attempting to naively predict a protein’s tertiary structure solely from its primary sequence is also intractable.

In simplest terms, the Scramble Squares objective is to place all of the given squares (tiles) into a 3 x 3 grid, such that every side (edge) shared between any two adjacent tiles contains both halves of a complete image. For example, in the image below, tile #0 and tile #1 can be placed directly next to each other because their east/west adjacent sides complete the combined image (head & body) of a Shar Pei puppy:



> Note: Tile half images along the outside edges of the final 3 x 3 solution matrix do NOT need to complement their corresponding tile half image on the other side of the matrix.

Our goal is to write a computer program that can find **all** valid layouts (legal solutions) to the 3x3 Scramble Squares given a description of the original tiles. The ideal program must run within *a few seconds on a single PC.*

This is a daunting task. It is hard enough for a human to find even just one valid solution given days or weeks of time. How can we therefore expect to instruct a computer to find *EVERY* solution in just seconds?

When confronted with a new challenge, scientists often begin by transcribing all of their initial questions & observations into a lab notebook. Before they can know any answers, they must first ask the **right** questions. Understanding only starts when you begin to realize there are always more intricate & hidden underlying relationships between seemingly independent observations than you could have possibly noticed at first glance.

The journey of discovery is the act of formulating better and better questions to get a deeper level of understanding. It is important to document these emerging patterns in a notebook for your future reference.

To give an example of the kinds of systematic observations which scientists might make when first presented with the challenge of analyzing & solving the Scramble Squares puzzle, please answer the following questions:

1. How many total squares (tiles) are there in your deck?
2. How many total distinct puppy breeds are represented in your deck?
3. How many half images (head or body) can be shown on one tile?
4. How many times does <u>each</u> image half, head vs. body, appear in your deck, for <u>each</u> breed?
5. Which tile(s) has the most number of head images?
6. Which tile(s) has the most number of body images?
7. Which breed has the highest number of complete (both head and body) images?
8. Which breed has the largest disparity (difference) between the numbers of head images vs. body images throughout your entire deck?

Now consider these parallels to protein folding:

- To be a valid placement, the body/head half images in the tiles must match up in complementary **pairs**. What are the base pairs in DNA? What are they for RNA?
- Each tile has four “binding” sites, one on each of the four sides of a tile. However, binding sites along the *outside* edges of the 3 x 3 matrix do not have to bind to any other tile edge. This is similar to how distant molecules in a protein strand do not normally influence each other as the protein folds into its final 3D structure. It is the molecules that are closest to each other that most directly affect the final fold.
- Various points along a protein the molecule can rotate around their bond angles as the strand reaches its final conformational shape. Each of these possible rotations represents an additional degree of freedom and this is one reason for the exponential growth in the number of possible configuration states a molecule might obtain during the folding process.

Given these analogies, answer these additional questions:

9. How many binding sites must be matched by the tile at the center of a valid 3 x 3 matrix?

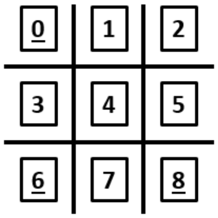
10. How many binding sites must be matched by the tiles at the corners of a valid 3 x 3 matrix?

11. How many ways can a tile rotate to present alternative binding sites to its neighboring tiles?

> Note: for purposes of this analysis, all rotations will be represented as an integral number of **counter-clockwise** quarter turns**.** So writing an "**r 1**" would mean a single quarter turn counter-clockwise, while an "**r 3**" would represent rotating a tile three (3) quarter turns counter-clockwise from its initial orientation.

A key aspect in understanding how to tackle the Scramble Squares puzzle is to differentiate the concept of a tile from a position. A "position" is a distinct location in the 3 x 3 matrix where any given tile could be placed, assuming all the complementary images halves at each of the requisite binding sites align to its adjacent neighboring tiles.

For purposes of the analysis, we shall adopt a convention where positon 0 is in the top-right corner of the matrix, increasing left to right, then top to bottom. The reason that all things in computer science (programming) start with the number zero will be discussed in class, but it is a common convention everyone follows. So the first position is position #0, not position #1:



0	1	2
3	4	5
6	7	8

> Note: The numbers zero, six, and eight will be underlined in this document to establish the canonical (meaning: proper, initial, or given) orientation. In the absence of the underline, it would be impossible to disambiguate the tile with the identification number of "8" that had either no rotation (by an **r 0**) or had been flipped upside down (by an **r 2**) since the number eight in Arabic numerals is symmetric around the horizontal axis.

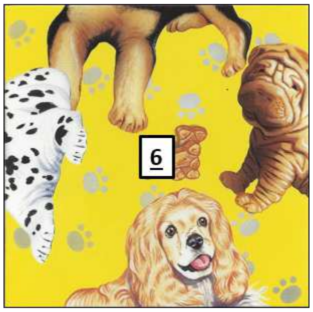
12. How many distinct positions are there in a 3 x 3 matrix?

13. What position #'s only need to match three binding sites in a valid solution?

One thing scientists often do early on during their investigations, in order to share their discoveries with other scientists, is to come up with a way to identify and classify things in a consistent fashion. They must develop and agree upon a “scheme” to explain their experiments and findings.

Given a deck of nine tiles, we must devise a numbering scheme to uniquely identify each tile. **The scheme can be wholly arbitrary, but it must be consistent**. Therefore, the original Puppy Scramble Squares puzzle tiles received from the manufacturer were assigned a random number between #0 and #8. Every student has the exact same set of tiles, numbered in the exact same sequence. This number also establishes the canonical (initial) orientation of the tile in terms of what half images appear in the north, south, east and west binding sites.

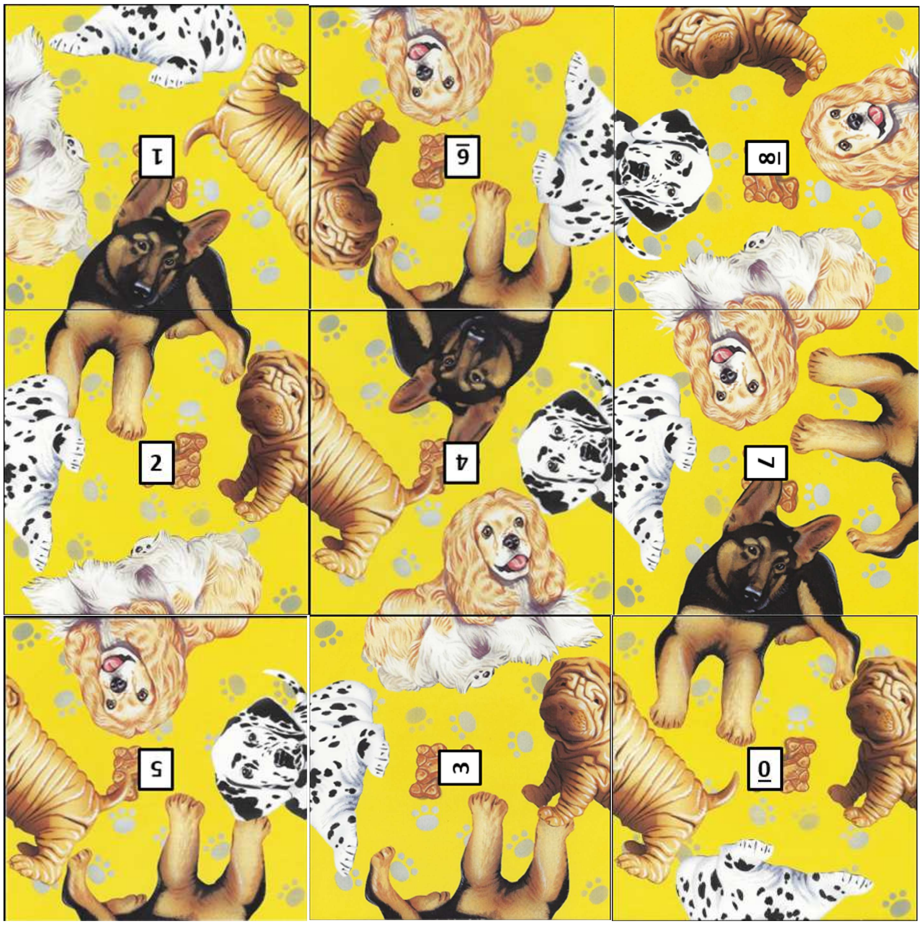
For example, shown below is tile #6 in its canonical orientation per our numbering scheme. The underline at the bottom of the number 6 (to disambiguate it from the number 9) establishes that in the canonical (**r 0**) orientation, tile #6 has a German Shepard body in the North binding site, a Shar Pei head in the East binding site, a Spaniel head in the South binding site, and Dalmatian body in the West binding site.



To check your understanding, here is one valid solution to the Puppies puzzle which was found by the computer program in less than 13 millionths of a second. Each parenthesis represents a position, with the first number being the tile #, and the second number being the number of counter-clockwise quarter turns that should be applied to that tile in that position:

(1 r 2)	**(6 r 2)**	**(8 r 1)**
(2 r 0)	**(4 r 2)**	**(7 r 3)**
(5 r 2)	**(3 r 1)**	**(0 r 0)**

This means that tile #1 is in position #0 (top right corner), and the tile is rotated two quarter turns counter-clockwise. Likewise, tile #0 would be placed in position #8 (the lower right corner of the matrix) and that tile would not be rotated at all from its canonical position. Recall the canonical position is where the number is upside right and the underline appears below the number. To check your understanding, please layout your tiles according to this solution and verify your final image looks like the following:



Notice the 12 inner binding sites are complimentary to each other, but not all of the 12 external (outer edge) bindings sites are complimentary to the half images on the opposite side of the matrix. A valid solution requires only that those inner 12 images match their adjacent tiles.

Given this new level of understanding – what strategy would you employ to find at least one valid solution to the puzzle? Specifically, which tile would you place first, where would you place it (into what position), and why? **What characteristics about each tile might guide you in determining the right sequence?**

After spending a few minutes in vain trying to "thoughtfully" place ALL of the tiles into a 3 x 3 grid, while obeying the rule that adjacent tile within the puzzle edges must complete an image (head-body or body-head, but not head-head or body-body), it is only human nature to give up.

Every layout starts out well enough, until it becomes impossible to find any remaining tile that works. Then you backup and try again. It seems impossible to discover any heuristics (rules of thumb) that could help guide you quickly to the solution. The frustrated natural reaction is then to just start trying every single possible permutation of layouts until you find one that works.

But is this brute force approach of trying every possible layout a tractable (realistic, obtainable) method for finding a solution? How long might it take to try every possible layout?

Answer the following questions:

14. With 9 tiles and 9 possible positions to place each tile, how many distinct ways can you lay them out into a 3 x 3 matrix? Hint: should you calculate permutations or combinations? Is order important when associating each tile # to a position #? Can you reuse a tile more than once in a given layout?
15. How many distinct ways can all 9 tiles be spun through each of the 4 possible rotations?
16. In total, how many possible ways can you layout the tiles, with all possible rotations, irrespective of whether the layout is valid or not according to the rules of Scramble Squares?

Consider how to answer question # 16:

- We must use <u>permutations</u> because order is important, and remember that each tile can only be used once in a given layout.
- Using Wolfram Alpha (http://www.wolframalpha.com) and entering “9 permute 9” returns **362,880**. So this means 9 tiles can be arranged into 9 distinct permutations in 362,880 ways.
- For each of these 362,880 permutations, each of the 9 tiles and be rotated in 4 different ways. So a tile in position #0 could be turned 4 different ways, multiplied by the 4 rotations of a tile in position #1, multiplied by the 4 rotations of a tile in position #2, etc. This means we have 4 * 4 * 4 * 4 * 4 * 4 * 4 * 4 * 4 = (4^9) = **262,144** distinct ways a given permutation of 9 tiles can be rotated.
- Therefore, the 362,880 permutations x 262,144 rotations (per each permutation) = **<u>95,126,814,720</u>** (95 billion!) total possible unique ways we can layout the tiles. Stop and consider how interesting it is that you can calculate ahead of time the total number of layouts even if you don’t know what they all are yet!

So how big of a number is 95 billion? If you could write a program so that a computer could evaluate a new layout every single millisecond, how long would it take to try them all? Answer: about **<u>3 years</u>**! [95,126,814,720 / 1000 / 60 / 60 / 24 / 365 = 3.0165] That is too long!

As scientists collaborate, they often propose alternative theories or even differing answers (computations) using the exact same theories. This is the nature of discovery: brilliant breakthroughs followed by long periods of trial & error and fumbling around in the dark until the next big breakthrough, and then the cycle repeats itself.

For example, consider what it would mean if a scientist proposes an alternative theory as to the "total number of unique Scramble Squares layouts" - given the very same 9 tiles used in our first theory:

- There are 9 distinct tiles. Each tile can be in one (and only one) of 4 possible rotations. Therefore, if you think about this, that means there are actually 9 x 4 = **36** distinct "tiles".
- 36 tiles taken 9 at a time (order is important) means 36 permute 9. Entering the text "36 permute 9" at http://www.wolframalpha.com shows 34,162,713,446,400
- Accordingly, there are actually **34,162,713,446,400** (34 trillion) total possible unique ways we can layout the tiles, which is *360* times more than our first theory predicted.

So now we have two different theories which predict two wildly different values for the total number of potential layouts. Note: whether a given layout is a valid or invalid solution does matter yet as we are counting only the total number of possible unique layouts.

17. Which theory is right? Could they both be right? Could both wrong?
18. What assumptions went into each theory?
19. How can one theory be off by a factor of *360*?
20. How could we prove which theory is correct?
21. What makes the naive brute force method so inefficient for solving these kinds of problems?

If the second theory is right, our brute force computer program will now require **1,080** years to try every possible layout. Yet consider that a typical protein can have over "150 thousand billion billion" (1.5 x 10^{23}) possible layouts, yet it finds its optimal layout (its own puzzle solution) in less than 3 milliseconds. Nature does not use brute force!

In class we will review the code to solve the Scramble Squares. This tool will allow us to perform additional "experiments" on the puzzle to uncover more hidden relationships between the tiles and images. However, to gain further insight into the true nature of the puzzle, please consider these final questions now, before we analyze the program:

22. How could you prove a set of given tiles has no solutions before even starting?
23. What makes one set of tiles harder to solve than other sets?
24. What group operations (substitutions, transpositions, rotations) can we use to exploit any hidden symmetries in a given tile set to find solutions even more quickly? Symmetry is a very powerful concept in nature, and plays a crucial role in quantum mechanics, which is the foundation of all biological chemistry. Please read this snippet: http://en.wikipedia.org/wiki/Symmetry#In_chemistry

As a hint to question #24, verify that following matrix is yet another solution to the Puppy Scramble Squares puzzle, given the numbering scheme we adopted for your tiles. Use the tile set provided and position the tiles according to this layout:

(8 r 2)	**(7 r 0)**	**(0 r 1)**
(6 r 3)	**(4 r 3)**	**(3 r 2)**
(1 r 3)	**(2 r 1)**	**(5 r 3)**

How does this second solution compare to the first solution provided?

(1 r 2)	**(6 r 2)**	**(8 r 1)**
(2 r 0)	**(4 r 2)**	**(7 r 3)**
(5 r 2)	**(3 r 1)**	**(0 r 0)**

Hint: Notice the middle tile (position #4, which also happens in this case to be tile #4) doesn't change, only its rotation. Follow the four corner tiles between both solutions – does the whole matrix itself appear to be rotating? What is happening to the number of the rotation for each tile?